Journal of the OPTICAL SOCIETY *of* AMERICA

VOLUME 38
No. 4

APRIL
1948

Photoelectric Method for Preparing Printing Plates . . . . . ARTHUR C. HARDY AND F. L. WURZBURG, JR. 295
Color Correction in Color Printing . . . . . ARTHUR C. HARDY AND F. L. WURZBURG, JR. 300
Electronic Method for Solving Simultaneous Equations . . . . . ARTHUR C. HARDY AND EDWARD C. DENCH 308
Absolute Sensitivity Measurements on Single-Grain-Layer Protographic Plates for Different Wave-Lengths . . . . . J. H. WEBB 312
Physics Abstracts 323
Performance of an Optical System as the Sum of the Contributions of Its Construction Elements . M. HERZBERGER 324
Visibility on Cathode-Ray Tube Screens: Signals on a P7 Screen Seen at Different Intervals after Excitation . . . . . A. L. SWEET AND N. R. BARTLETT 329
International Commission of Optics 337
Phase Microscopy with Vertical Illumination . . . . . HELEN JUPNICK, HAROLD OSTERBERG, AND G. E. PRIDE 338
Centering of Optical Systems . . . . . A. L. INGALLS AND K. PESTRECOV 343
New Determination of the Young-Helmholtz Primaries . . . . . GEORGE SZEKERES 350
Visibility on Cathode-Ray Tube Screens: The Effects of Light Adaptation . . . R. M. HANES AND S. B. WILLIAMS 363
Investigation of the Use of Geiger-Müller Photoelectron Counters in the Spectrographic Analysis of Phosphorus in Steel . . . . . RICHARD HANAU AND RALPH A. WOLFE 377
Effects of Illumination Level on Visual Performance and Fatigue . . . . . ERNST SIMONSON AND JOSEF BROZEK 384
Photographic Granularity and Graininess. IV. Visual Acuity Thresholds; Dynamic *versus* Static Assumptions . . . . . LOYD A. JONES AND GEORGE C. HIGGINS 398
Color of Unstained Wood . . . . . PARRY MOON AND DOMINA EBERLE SPENCER 405
Centennial Celebration of A.A.A.S. 408
Electroformed Precision Prisms . . . . . BENJAMIN GOLDBERG 409
Behavior of an Interferometer in a Gravitational Field. II. Application to a Planetary Orbit . . . HERBERT E. IVES 413
Book Review 328
Letters to the Editor:
"Optical Staining" of Tissue . . . . . GERMAIN C. CROSSMON 417
Note on Ultraviolet Absorption of Zinc Oxide Phosphors . . . . . F. H. NICOLL 417
Nomographs for Spectral Band Width of the Beckman DU Spectrophotometer . . H. M. HAENDLER 417
Electro- and Magneto-Optical Shutters for Aerial Cameras . . . . . ROY C. GUNTER, JR. 419

Published monthly
at Prince and Lemon Sts., Lancaster, Pa.
by the
AMERICAN INSTITUTE OF PHYSICS
Incorporated
57 East 55 Street, New York 22, New York
FOR AND UNDER THE EDITORIAL SUPERVISION OF THE
OPTICAL SOCIETY OF AMERICA

THE OPTICAL SOCIETY OF AMERICA

Officers of the Society

RUDOLF KINGSLAKE, *President*
Eastman Kodak Company
Rochester, New York

WILLIAM F. MEGGERS, *Vice President*
National Bureau of Standards
Washington 25, D. C.

ARTHUR C. HARDY, *Secretary*
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

W. W. GRAEPER, *Treasurer*
Bausch & Lomb Optical Company
Rochester 2, New York

The Board of Directors consists of the above officers; the Editor of the Journal; the Past President; and five directors-at-large.

G. R. HARRISON, *Editor and Past President*
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

Directors-at-Large

FRANK BENFORD
General Electric Company
Schenectady, New York

F. M. BISHOP
Eastman Kodak Company
Rochester 4, New York

WALLACE R. BRODE
Ohio State University
Columbus 10, Ohio

A. H. BENNETT
American Optical Company
Buffalo 11, New York

E. H. LAND
Polaroid Corporation
Cambridge, Massachusetts

Local Sections

STANLEY S. BALLARD, *Secretary for Local Sections*

Rochester Section

ARTHUR F. TURNER, *President*
ROBERT N. WOLFE, *Vice President*
PAUL W. STEVENS, *Secretary*
HOMER C. IMES, *Treasurer*

Detroit Section

GERALD M. RASSWEILER, *Chairman*
JOHN SCHUCH, *Vice Chairman*
DANIEL L. HARMON, *Secretary*
HENRY TUTTLE, *Treasurer*

Niagara Frontier Section

ALLEN B. FUHR, *President*
GUY H. FETTERLEY, *Secretary*
HOYT G. THOMPSON, *Treasurer*

Patrons

AMERICAN OPTICAL COMPANY — BAUSCH & LOMB OPTICAL COMPANY
EASTMAN KODAK COMPANY

Corporation Members

ALLIED CHEMICAL AND DYE CORPORATION, NATIONAL ANILINE DIVISION, Buffalo, New York
AMERICAN CYSTOSCOPE MAKERS, INC., New York, New York
AMERICAN OPTICAL COMPANY, Southbridge, Massachusetts
ARGUS, INCORPORATED, Ann Arbor, Michigan
BAUSCH & LOMB OPTICAL COMPANY, Rochester, New York
BELL & HOWELL COMPANY, Chicago, Illinois
BELL TELEPHONE LABORATORIES, INC., New York, New York
CENTURY LIGHTING, INC., New York, New York
CORNING GLASS WORKS, Corning, New York
R. R. DONNELLEY & SONS COMPANY, Chicago, Illinois
E. I. DUPONT DE NEMOURS & COMPANY, INC., Chemical Department, Experimental Station, Wilmington, Delaware
EASTMAN KODAK COMPANY, Rochester, New York
EDMUND SALVAGE COMPANY, Oaklyn-Audubon, New Jersey
FAIRCHILD AVIATION CORPORATION, Jamaica, New York
FARRAND OPTICAL COMPANY, INC., New York, New York
FISH-SCHURMAN CORPORATION, New York, New York
GAERTNER SCIENTIFIC CORPORATION, Chicago, Illinois
W. & L. E. GURLEY, Troy, New York
INTERCHEMICAL CORPORATION, New York, New York
KEUFFEL & ESSER COMPANY, Hoboken, New Jersey
KOLLSMAN INSTRUMENT DIVISION OF SQUARE D COMPANY, Elmhurst, New York
NATIONAL CARBON COMPANY, Cleveland, Ohio
NATIONAL TECHNICAL LABORATORIES, South Pasadena, California
PERKIN-ELMER CORPORATION, Glenbrook, Connecticut
POLAROID CORPORATION, Cambridge, Massachusetts
PORTMAN MACHINE TOOL COMPANY, New Rochelle, New York
SIMPSON OPTICAL MANUFACTURING COMPANY, Chicago, Illinois
SOCONY-VACUUM LABORATORIES, Paulsboro, New Jersey
ARTHUR H. THOMAS COMPANY, Philadelphia, Pennsylvania
UNIVIS LENS COMPANY, Dayton, Ohio
ZENITH OPTICAL COMPANY, Huntington, West Virginia

THE OPTICAL SOCIETY OF AMERICA was organized in 1916 "*to increase and diffuse the knowledge of optics in all its branches, pure and applied, to promote the mutual interests of investigators of optical problems, of designers, manufacturers and users of optical instruments and apparatus of all kinds and to encourage cooperation among them.*" It aims to serve the interests of all who are engaged in any branch of optics from fundamental research to the manufacture of optical goods.

MEMBERSHIP

Regular Members: "Any person who has contributed materially to the advancement of optics is eligible for regular membership."

Associate Members: "Any person who is interested in optics is eligible for associate membership."

Corporate Members: "Any corporation desirous of encouraging the work of the Society is eligible for corporation membership."

Patrons: The payment, within any one year, of $1000 or more toward the expenses of the Society by any corporation or individual entitles such corporation or individual to the rank of patron.

The Society cordially invites to membership all who are interested in any branch of optics, either in research, in instruction, in optical or illuminating engineering, in the manufacture and distribution of optical goods of all kinds, or in physiological and medical optics.

ANNUAL DUES

Regular members, $8.50; Associate members, $6.00; Corporate members, $50.00. These dues include subscription to the Society's official journal, the *Journal of the Optical Society of America.*

CHANGES OF ADDRESS

All changes of address of members of the Optical Society of America should be communicated to the Secretary, Professor Arthur C. Hardy, Massachusetts Institute of Technology, Cambridge 39, Massachusetts.

Journal of the
OPTICAL SOCIETY
of AMERICA

VOLUME 38, NUMBER 4 APRIL, 1948

A Photoelectric Method for Preparing Printing Plates[1]

ARTHUR C. HARDY
Massachusetts Institute of Technology, Cambridge, Massachusetts

AND

F. L. WURZBURG, JR.
Interchemical Corporation, New York, New York

(Received December 22, 1947)

Whenever the subject being reproduced has a regular pattern, the conventional type of halftone screen, because of its regular structure, may produce an objectionable beat pattern, called *moiré*. In multicolor printing, the regular patterns of the several structured images produce complex beat patterns whose effects are minimized, but not eliminated, through proper choice of screen angles. This paper describes a procedure in which a continuous-tone image is scanned; and the resulting photo-tube signal is used to actuate a light valve in such a manner as to produce a structured image containing dots that are, in effect, distributed at random. It is shown that this procedure avoids both the distortions in tone reproduction and the moiré patterns associated with the conventional use of halftone screens.

AS is well known, most printing processes used in the graphic arts require that continuous-tone images be translated into structured images having dots of variable area. The purpose of this paper is to describe a method designed to avoid some of the shortcomings of the conventional type of structured image, especially in color printing.

The conventional type of structure is represented, greatly magnified, in Fig. 1 for dot ratios of 0.1, 0.5, and 0.9. By definition, the dot ratio is the ratio of printed area to the total area, and it ordinarily depends on the exposure received by a photographic material located a short distance behind a halftone screen in the engraver's camera. Each of the openings in the halftone screen forms a "pinhole" image of the lens diaphragm; but, by using a photographic material of high contrast, these diffuse images are converted into opaque dots, the size of each being dependent upon the illumination of the corresponding opening in the screen.

H. E. Ives[2] has shown that this process results in serious tone distortion unless more elaborate techniques are employed than is normally the case. Furthermore, although the dots appear on casual inspection to have sharp edges, there is a pronounced variation in density across each dot. This variation in density causes the dots on the resulting printing plate to vary in size with variations both in exposure of the resist and in

[1] Extension of paper presented at the October, 1946 meeting of the O.S.A.

[2] H. E. Ives, "Tone reproduction in the 'halftone' process," J. Opt. Soc. Am. and Rev. Sci. Inst. **13**, 537 (1926).

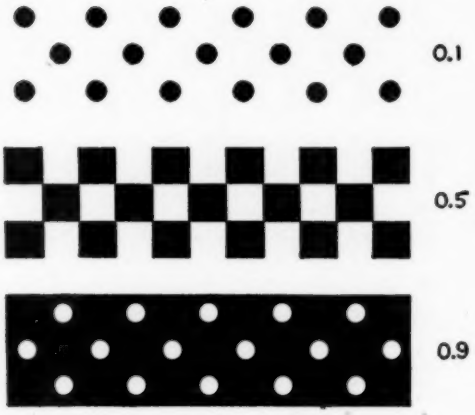


FIG. 1. Conventional type of structured image for dot ratios of 0.1, 0.5, and 0.9.

the subsequent etching of the plate. These difficulties account, at least in part, for the extensive amount of re-etching by hand that is ordinarily required, even for black and white reproductions.

The conventional structured image, because of its regularity, often exhibits a "beat" pattern, called moiré. As an example, if the subject happens to depict a shingled roof or a brick wall, the regular pattern of the subject may form a conspicuous interference figure with the pattern of the structured image. In color printing, the superposition of one regular dot pattern upon another is, in itself, a potential source of moiré. Ideally, printing plates made from the same halftone screen and printed with their dots exactly superimposed would exhibit no interference pattern. This ideal is unrealizable in practice because a slight shrinkage of the paper between impressions or a slight misorientation of one impression with respect to another would lead to a very objectionable moiré pattern. Even if these factors could be controlled, a slight shift in the register, amounting to no more than half the distance between adjacent dots, would considerably alter the color reproduction because, even with the best inks available, the additive mixture produced by two dots side by side does not give the same average color as the subtractive mixture when they are superposed.

For the above reasons, the halftone screen is invariably rotated through a sizable angle in its own plane after each negative of a color set has been exposed. By thus deliberately introducing a comparatively fine moiré pattern, the coarse moiré pattern that would result from shrinkage of the paper or faulty registration is avoided. Also, when the screen is thus rotated, the average color of an area that is large by comparison with the size of a dot does not depend upon the registration of the successive impressions but is, in fact, the same as though the dots of each plate were randomly distributed.

Experience has shown that rotating the screen through 30° after each exposure produces the least objectionable beat pattern, and this has long been standard practice in three-color printing. In four-color printing, however, such a rotation would bring the fourth plate into a position at right angles to the first; and, because of the symmetry of the screen, this is the equivalent of no rotation. Since a $22\frac{1}{2}°$ difference in screen angles produces a very objectionable pattern, it is common practice in four-color printing to orient the screens for the black, magenta, and cyan impressions so that they are separated by 30°, and to orient the yellow screen so that it lies halfway between two of the others, usually between the magenta and cyan. Although the moiré pattern formed by two impressions with screen angles separated by only 15° is ordinarily very pronounced, the fact that there is less difference between the additive and subtractive mixtures of pairs of colors which include yellow as one of the components makes this the best solution for general use in four-color printing. Nevertheless, the moiré pattern may be very conspicuous, particularly in certain flesh tones.

In view of these undesirable features of the conventional halftone structure, our first experiments were designed to investigate dot structures that would be inherently free from moiré patterns. It will be obvious from the foregoing that the single-line screens that were used to some extent before the development of cross-line screens would present no difficulty in four-color printing from the standpoint of moiré pattern, provided the directions of the lines in the successive impressions make appreciable angles with one another. Of course, structured images produced by single-line screens have a limited tone range under practical printing conditions, but it seemed to us that this limitation might be re-

moved if the lines were broken up into dots whose spacing could be appropriately varied. These considerations suggested photoelectric scanning, which had previously been used in facsimile transmission by wire and radio. We felt that the cylindrical scanning customarily employed in these fields should be replaced by flat scanning, partly to adapt the procedure to the use of photographic plates as well as films, but more because of the necessity of scanning the several plates of a color set at different angles.

A machine adapted for flat scanning is shown schematically in Fig. 2. Normally the photo-tube scans a continuous-tone positive transparency made from the original subject. This photographic step seems desirable because the original is seldom of the correct size for reproduction. A positive is preferred to a negative in order to satisfy the requirements of faithful tone reproduction. By imaging a slit like that shown in Fig. 3 onto an unexposed photographic material on the receiving platen, a dot occupying the full width of the scanning line is generated whenever the slit is illuminated. One convenient way of controlling the illumination is to use a mirror type of light valve to throw a beam of light across the slit whenever the light valve is supplied with a suitable signal (Fig. 3a). In the absence of such a signal, no light reaches the slit and a space is generated (Fig. 3b). With a sufficiently narrow slit, this procedure results in dots that are uniform in density across their surfaces.

In the electrical circuit employed, the photo-tube current (after amplification) charges a con-

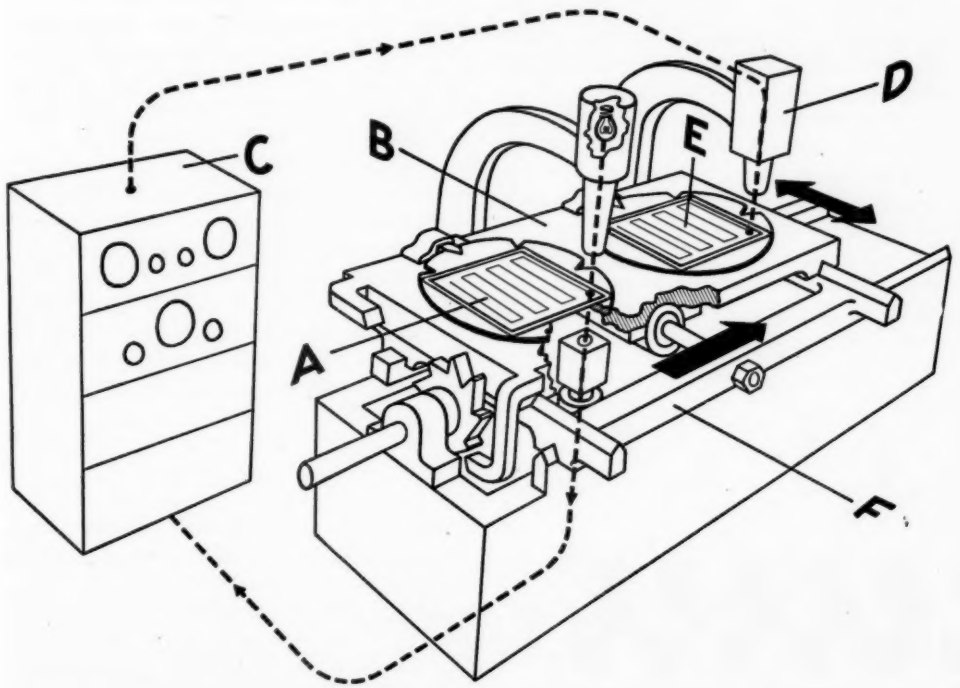


FIG. 2. Schematic representation of the operation of a machine adapted for flat scanning. A continuous-tone positive *A* is attached to a circular platen which is clamped to the carriage *B*. An hydraulic piston imparts a reciprocating motion to this carriage as shown by the double-ended arrow. At the end of each stroke a lead screw and ratchet mechanism causes the intermediate carriage *F* to advance a short distance in the direction of the single-ended arrow. Light that passes through this transparency during the scanning operation falls on the cathode of a photo-tube, and the resulting signal controls the dot-producing circuit in the relay rack *C*. The output of the relay rack actuates a light valve in the recording head *D*, producing an intermittent exposure that results in recording a structured image on the unexposed photographic plate *E*. When used for multicolor reproduction, the circular platens are rotated through an appropriate angle after each recording and are held in position by means of the detents shown.

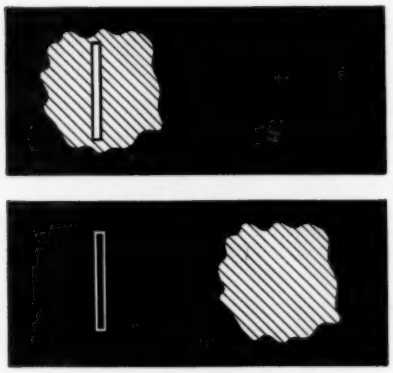


FIG. 3. In producing the type of structured image illustrated in Fig. 4, the slit is illuminated by a beam of light reflected from the galvanometer mirror when generating a dot (Fig. 3a) and receives no light when generating a space (Fig. 3b).

denser to a predetermined trigger value. At this point, the light valve suddenly starts the generation of a dot, and simultaneously a constant current device (acting in opposition to the photo-tube) starts to discharge the condenser. When the condenser has been discharged to a somewhat lower trigger value, the generation of the dot is suddenly terminated and the constant current device is rendered inoperative.

The time t_c required for the photo-tube current i to charge the condenser from the lower trigger value to the higher is

$$t_c = k/i, \tag{1}$$

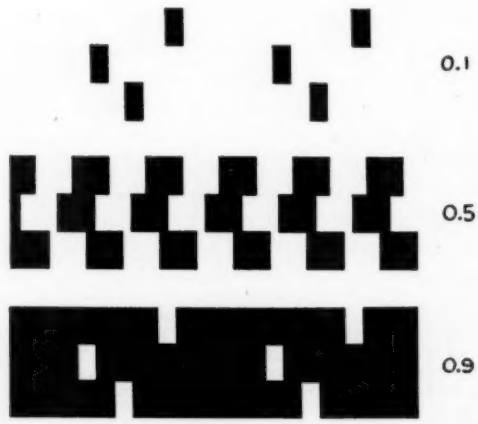


FIG. 4. Type of structured image produced by the method illustrated in Fig. 3.

where k is a proportionality constant. The time of discharge t_d from the higher trigger value to the lower depends upon the difference between the constant discharge current I and the photo-tube current i. Thus,

$$t_d = k/(I-i). \tag{2}$$

Since the length of a dot is proportional to t_d and the space between the dots is proportional to t_c, the ratio of printed area to total area is $t_d/(t_d+t_c)$. This is the dot ratio r, which is found by substitution to be

$$r = i/I. \tag{3}$$

It will be seen from this equation that, by adjusting the discharge current I until it is equal to the photo-tube current produced when scanning a white area of the original, the dot ratio becomes unity in the white areas and zero in the black areas. It will be seen, also, that the requirements of faithful tone reproduction are fulfilled at all intermediate tone values. This type of structure is represented, greatly magnified, in Fig. 4 for dot ratios of 0.1, 0.5, and 0.9.

One of the disadvantages of the structure produced by the method described is a tendency for the dots in adjacent lines to "line up," especially in the vicinity of a sharp boundary between two areas of different tone value. This tendency can be largely overcome by automatically altering the time constant of the dot-producing circuit by a small amount between adjacent scanning lines. The cycling frequency f of the circuit is

$$f = \frac{1}{t_c+t_d} = \frac{r(1-r)}{k}, \tag{4}$$

where the constant k depends upon the time constant of the circuit. Thus, altering the time constant changes the frequency and hence the number of dots in a given length of the scanning line. Since the constant k does not appear in Eq. (3), the tone reproduction is unaffected.

Another disadvantage of this type of structure is apparent from Eq. (4). If the constants of the dot-producing circuit are adjusted until the number of dots per inch of scanning line is satisfactory in the middle tones, where the frequency is at a maximum, the number of dots per inch may be insufficient in the lightest and darkest tones.

Indeed, Eq. (4) indicates that the frequency approaches zero as the dot ratio r approaches either zero or unity. This extreme variation in frequency can be reduced by actuating the dot-producing circuit, not by the photo-tube current itself, but by some appropriate function thereof. For example, if the dot-producing circuit is actuated by a current i', where

$$i' = ai + I', \qquad (5)$$

i' will never become zero because of the constant current I' and will never become unity if the constant a is properly chosen. Since this reduction in the frequency range results in a closer spacing of the dots in the lighter tones and a closer spacing of the spaces between the dots in the darker tones, the rendering of tone values would be altered unless some form of compensation were introduced, *viz.*, appropriate variation in the widths of the dots and spaces.

We have found that the widths of both the dots and the spaces can be varied by imaging an aperture of appropriate shape on the slit. In Fig. 5 the shaded area represents the image of such an aperture in the plane of the slit, this image replacing the irregularly shaped light patch of Fig. 3. An additional coil of the galvanometer is supplied by a biasing signal proportional to the photo-tube current which shifts the positions of the image of the aperture with respect to the slit. Thus, when scanning a dark area of the original, the photo-tube current is small and the image of the aperture oscillates between the two positions shown at **a** and **b** in Fig. 5. When scanning a light area of the original, the image of the aperture oscillates between the positions shown at **c** and **d** in Fig. 5. This results in a dot structure like that represented in Fig. 6.

It will be evident that the shape of the aperture is determined by the choice of the constants a, I, and I'. There are certain restrictions applying to the choice of these constants because the width of a dot or space can never be less than zero nor greater than the width of the scanning line. If it is decided to make t_e equal to t_d in the middle tone where the dot ratio is 0.5 and to cause the frequency to decrease symmetrically on either side of this middle tone value, the constants are

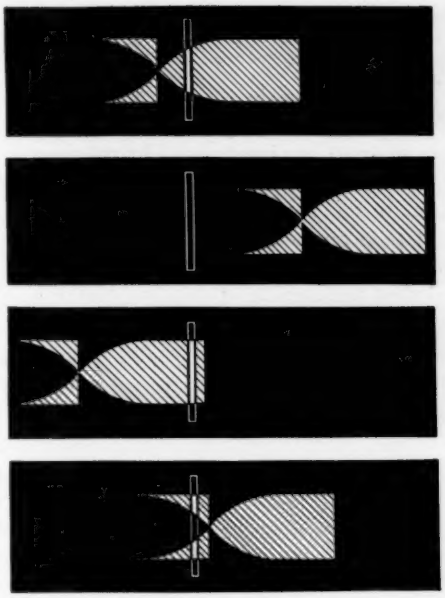


FIG. 5. An extension of the method of Fig. 3 used in producing the type of structured image illustrated in Fig. 6. As before, the image of the aperture is in position **a** when generating a dot and in position **b** when generating a space, the width of a dot being determined by the location of the image relative to the slit in position **a**. A second coil of the galvanometer is supplied with a current proportional to the photo-tube signal and this moves the image of the aperture to the left as the photo-tube signal increases. The shape of the aperture is such that dots of increasing width are produced until the width of the dot equals that of the scanning line. Further increase in the photo-tube signal results in the production of spaces of decreasing width. In this illustration, positions **a** and **b** represent the formation of an image with a dot ratio of 0.1, and positions **c** and **d** represent the formation of an image of dot ratio 0.9.

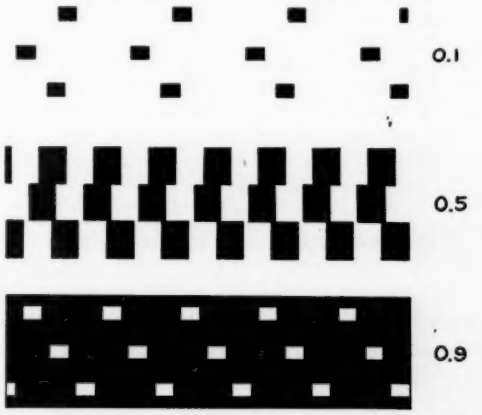


FIG. 6. Type of structured image produced by the method illustrated in Fig. 5. This diagram does not illustrate the change in frequency that is automatically introduced after each line has been scanned.

related in the following manner

$$a = I - 2I'. \tag{6}$$

The aperture shown in Fig. 5 was designed to satisfy this relationship.

This program was interrupted by the war at a time when we had made only a few printing plates by this method. Our tests have shown that faithful tone reproduction is easily achieved in the photographic images; but it should be mentioned that the etching of printing plates is, in general, a non-linear process. This non-linearity can be compensated, of course, by making appropriate modifications in the shape of the aperture used to vary the width of the dots and spaces. As might be expected, four-color reproductions having this quasi single-line pattern did not exhibit the moiré that is characteristic of conventional process printing. It will be shown in the following paper that full color correction can be automatically introduced during the scanning operation.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 38, NUMBER 4 APRIL, 1948

Color Correction in Color Printing[1]

ARTHUR C. HARDY
Massachusetts Institute of Technology, Cambridge, Massachusetts

AND

F. L. WURZBURG, JR.
Interchemical Corporation, New York, New York

(Received December 22, 1947)

In a three-color print, the color of any area that is large in comparison with the size of the dots of the structured image can be regarded as an additive mixture of eight colors: the unprinted paper stock; the cyan, magenta, and yellow of the individual ink dots; the red, green, and blue that result when ink dots overlap in pairs; and the black that results when all three ink dots overlap. Since the extent of overlapping is determined by the sizes of the dots, the color of the additive mixture can be expressed by three equations. The straightforward attack on the problem of color reproduction is to solve these fundamental equations for the required dot sizes on the assumption that, in every area, the tristimulus values of the reproduction should be equal to those of the original. This has not been accomplished hitherto, because it involves solving three simultaneous equations of third degree, each containing eight terms. However, a relatively simple electronic equation-solving network has been constructed which solves these equations with ample precision in 0.001 second. By using this network in connection with a scanning machine of the type described in the preceding paper, full color correction is achieved. A corresponding set of equations can be written for the additive mixture produced in four-color printing, wherein the fourth color is black. Since the three equations now contain four unknowns, an additional condition must be imposed. From the standpoint of the printing requirements, it is desirable that at least one of the color dots be absent, or of some predetermined minimal size, in every region of the reproduction. An extension of the principles embodied in the electronic network mentioned above imposes this condition and yields a continuous solution to the three fourth-degree equations.

THE preceding paper[2] has described a scanning machine that converts a continuous-tone photographic image into the structured type of image that is required by many printing processes. One purpose of this paper is to show how full color correction in either a three-color or a four-color process can be automatically introduced during the scanning operation.

STATEMENT OF THE PROBLEM

The starting point in any method of color printing is the making of three color-separation negatives, so-called. These are ordinarily made by photographing the subject (which may be a color photograph, an artist's sketch, or a scene

[1] Extension of paper presented at the October, 1946 meeting of the O.S.A.

[2] A. C. Hardy and F. L. Wurzburg, Jr., J. Opt. Soc. Am. 16, 295 (1948).

in nature) through a red filter, a green filter, and a blue filter. Regardless of the subsequent procedures, these color-separation negatives must properly evaluate the color[3] of every elemental area of the subject. Photographic positives made from these negatives will then represent the tristimulus values of every elemental area of the subject; and the representation will be linear if the photographic operations fulfill the usual requirements of faithful tone reproduction in monochrome (i.e., reproduce a gray scale correctly). If three such positives are scanned simultaneously in proper register on a three-platen scanning machine like that represented schematically in Fig. 1, the three photo-tube currents will indicate the tristimulus values of the subject as each elemental area is scanned in sequence. The problem is to use this intelligence to produce color-corrected structured images from which printing plates can be made by standard procedures, without resort to hand correction at any stage.

DOT RATIOS IN THREE-COLOR PRINTING

The fundamental problem set forth in the preceding section may be restated more precisely as follows. Let the tristimulus values of an elemental area of the subject be represented in some colorimetric system by R, G, and B. If the photographic positives on the scanning machine have been properly prepared, the three photo-tube currents may likewise be represented (in appropriate units) by R, G, and B. Then, if the tristimulus values of the corresponding area in the reproduction are represented by R', G', and B', the color reproduction will be perfect if

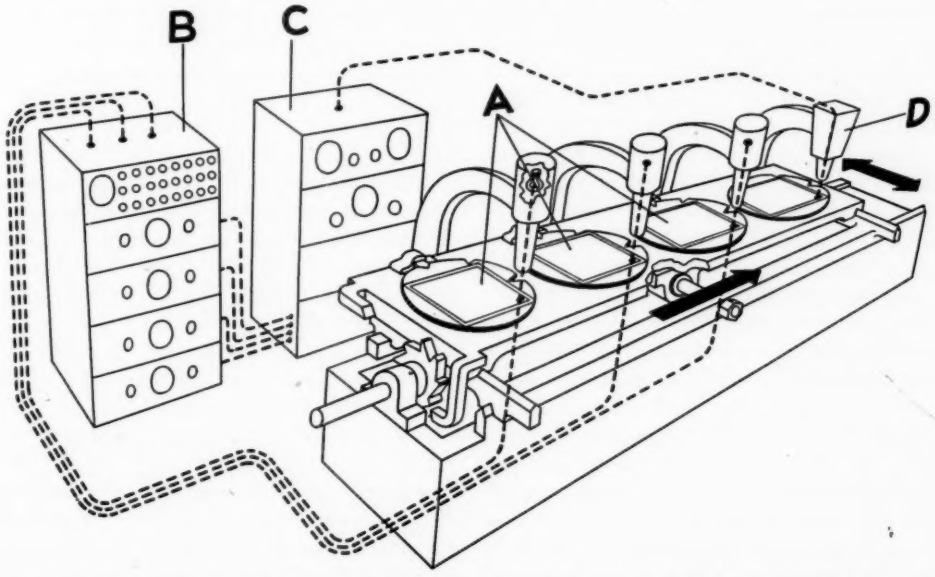


FIG. 1. Schematic representation of a scanning machine adapted for scanning three photographic positives simultaneously. The three positives *A* are attached to circular platens in the proper register. The platens are clamped to a carriage whose motions are identical with those described in connection with Fig. 2 of the preceding paper. Three scanning beams produce photo-tube signals which are fed to computing channels in relay rack *B* whose output signals control the dot-producing circuit in relay rack *C*. The output signal from the latter actuates a light valve in the recording head *D* producing an intermittent exposure that results in recording a structured image on an unexposed photographic plate clamped to the fourth platen. The circular platens are rotated through an appropriate angle after each recording.

[3] Since there may be areas of the subject which are unlike in spectral composition but nevertheless appear to the eye to be of the same color, it is evident that the color-separation negatives must evaluate these areas in the same manner as does the human eye. In other words, the gammas of the three emulsions must be independent of wave-length within the spectral range transmitted by each filter and the spectral sensitivity curves of the three filter-emulsion combinations must correspond to the color-mixture curves of a normal observer for some set of basic stimuli. In the cycle of operations to be described below, these are the only requirements that need be imposed on the filter-emulsion combinations used in making the color-separation negatives. In other words, the spectral sensitivities of the filter-emulsion combinations are quite independent of the colors of the printing inks.

TABLE I.

Color	Symbol	Fractional Area
White	w	$(1-c)(1-m)(1-y)$
Cyan	c	$c(1-m)(1-y)$
Magenta	m	$m(1-c)(1-y)$
Yellow	y	$y(1-c)(1-m)$
Red	my	$my(1-c)$
Green	cy	$cy(1-m)$
Blue	cm	$cm(1-y)$
Black	cmy	cmy

$R'=R$, $G'=G$, and $B'=B$ for every elemental area of the reproduction.

When printing from structured images, the tristimulus values R', G' and B' of any elemental area of the reproduction are controlled by the sizes of the printed dots that lie within this area. In three-color printing each elemental area contains only eight colors: the unprinted paper stock; the cyan, magenta, and yellow of the individual ink dots; the red, green, and blue that result when ink dots overlap in pairs; and the black that results when all three ink dots overlap. Practices vary widely with respect to the order in which the three inks are printed; but, since the theory to be developed here is applicable to any sequence, the alphabetical order (cyan, magenta, yellow) will be assumed. The dot ratios in any elemental area of the structured images will be represented by m, c, and y, respectively.

A sheet of white paper that has received only the cyan impreession will have dots of cyan ink covering a fraction c of each elemental area; whereas a fractional area $(1-c)$ will remain unprinted. When the magenta impression is superimposed, the magenta dots will overlap the cyan dots in a fractional area cm, and will fall on white paper in a fractional area $m(1-c)$, by virtue of the effectively random distribution of the dots. By the same reasoning, the fractional areas of the cyan dots and of the white paper remaining after the second impression are, respectively, $c(1-m$ and $(1-c)(1-m)$. When this reasoning is extended to include the third impression, the fractional areas occupied by the eight colors are found to be as given in Table I. The sum of these fractional areas is unity, of course.

In some colorimetric system, let the tristimulus values of the white paper be represented by R_w, G_w, and B_w. Similarly, let the tristimulus values of the other seven colors be represented by R, G, and B with appropriate subscripts from column 2 of Table I. Then, in terms of these constants, the tristimulus values R', G', and B' of the additive mixture of the eight colors within any elemental area are

$$\begin{aligned}R' = {} & (1-c)(1-m)(1-y)R_w \\ & +c(1-m)(1-y)R_c+m(1-c)(1-y)R_m \\ & +y(1-c)(1-m)R_y+my(1-c)R_{my} \\ & +cy(1-m)R_{cy}+cm(1-y)R_{cm} \\ & +cmyR_{cmy}, \end{aligned} \quad (1a)$$

$$\begin{aligned}G' = {} & (1-c)(1-m)(1-y)G_w \\ & +c(1-m)(1-y)G_c+m(1-c)(1-y)G_m \\ & +y(1-c)(1-m)G_y+my(1-c)G_{my} \\ & +cy(1-m)G_{cy}+cm(1-y)G_{cm} \\ & +cmyG_{cmy}, \end{aligned} \quad (1b)$$

$$\begin{aligned}B' = {} & (1-c)(1-m)(1-y)B_w \\ & +c(1-m)(1-y)B_c+m(1-c)(1-y)B_m \\ & +y(1-c)(1-m)B_y+my(1-c)B_{my} \\ & +cy(1-m)B_{cy}+cm(1-y)B_{cm} \\ & +cmyB_{cmy}. \end{aligned} \quad (1c)$$

It is obvious that these equations will accurately describe the color of the mixture of the eight components under ideal printing conditions; and Neugebauer,[4] who first published these equations, has shown that they apply in practice. The equations are therefore useful in predicting the color gamut that can be achieved with a given set of inks; but the primary purpose of this paper is to show how these equations provide a basis for automatic color correction.

It has been stated previously that perfect color reproduction will result when $R'=R$, $G'=G$, and $B'=B$ for every elemental area. This assumption of perfect reproduction can be introduced into the above equations by merely substituting R, G, and B for R', G', and B'. Then, if the new equations could be made explicit for c, m, and y, the proper dot ratios would be known in terms of the color of the subject and the constant colors of the eight components of the reproduction. The following section will describe a method by which values of c, m, and y which satisfy these equations can be determined and recorded as dots on structured images at a rate of more than one thousand dots per second.

[4] Neugebauer, Zeits. f. tech. Physik **36**, 22 (1937).

SOLUTION OF THE NEUGEBAUER EQUATIONS

It will be conducive to a better understanding of the method of solving the Neugebauer equations if numerical values are substituted for the constants, as has been done in Eqs. (2) for a set of high quality process inks and a typical printing routine. These constants were determined by standard procedures, the spectrophotometric characteristics of the eight colors being determined with a sheet of white paper as the white standard. These spectrophotometric data were then used to compute the tristimulus values of the eight colors, assuming Illuminant C and the standard observer.[5] On this basis,[6] the tristimulus values of the white paper are 1.000, 1.000, 1.000; those of the cyan ink are 0.146, 0.197, 0.613, etc.

$$\begin{aligned} R=R'=&1.000(1-c)(1-m)(1-y)\\ &+0.146c(1-m)(1-y)\\ &+0.442m(1-c)(1-y)\\ &+0.857y(1-c)(1-m)\\ &+0.398my(1-c)+0.064cy(1-m)\\ &+0.033cm(1-y)+0.023cmy, \end{aligned} \tag{2a}$$

$$\begin{aligned} G=G'=&1.000(1-c)(1-m)(1-y)\\ &+0.197c(1-m)(1-y)\\ &+0.219m(1-c)(1-y)\\ &+0.980y(1-c)(1-m)\\ &+0.206my(1-c)+0.196cy(1-m)\\ &+0.014cm(1-y)+0.019cmy, \end{aligned} \tag{2b}$$

$$\begin{aligned} B=B'=&1.000(1-c)(1-m)(1-y)\\ &+0.613c(1-m)(1-y)\\ &+0.185m(1-c)(1-y)\\ &+0.148y(1-c)(1-m)\\ &+0.012my(1-c)+0.103cy(1-m)\\ &+0.120cm(1-y)+0.016cmy. \end{aligned} \tag{2c}$$

It will be evident from inspection of the above equations that, if they were solved explicitly for c, m, and y by analytical procedures, the resulting expressions would contain terms of ninth degree. Doubting that such an analytical approach would be profitable, we[7] undertook to devise a means for handling the equations without making them explicit for c, m, and y. In view of the number of solutions required (one for each dot in the reproduction), it appeared that only electronic circuits could be made to operate with sufficient rapidity. The underlying principles involved in solving the equations electronically will be outlined here.

Imagine three electrical networks so constructed that, when supplied with input signals c, m, and y, their output signals will be R', G', and B', respectively. In other words, let each network perform the algebraic operations indicated by one of the equations. Imagine also three vacuum tubes whose plate currents provide input signals c, m, and y to these networks. The problem is then to adjust the values of these three plate currents until the output signals R', G', and B' are equal, respectively, to the phototube signals R, G, and B. In general, R', for example, will not be equal to R, but the disparity between R' and R can be used to control the grid of the vacuum tube whose plate current is c in such a manner as to tend always to reduce the disparity. Then, if the disparity between G' and G is used to control m, and the disparity between B' and B is used to control y, all three disparities can be reduced simultaneously. By introducing sufficient gain into these feed-back circuits, the three disparities can be made substantially zero. Under these conditions, the plate currents c, m, and y of the three vacuum tubes represent the required solutions of the equations. These currents can then be used, one at a time, to control the dot-producing circuit described in the preceding paper.

THE SCANNING MACHINE

In 1937, we published a paper[8] which indicated the nature of the corrections that are necessary in additive methods of color reproduction and showed that the same corrections

[5] American Standards Association, "Specification and description of color," Z44—1942.

[6] The assumption of standard colorimetric procedures is made in the interest of clarity of presentation, inasmuch as the meaning of tristimulus values in the I.C.I. system is now widely understood. By adopting this colorimetric system, however, freedom of selection of filter-emulsion combinations for the color-separation negatives is restricted to a single set; namely, a set whose spectral sensitivity curves correspond to the mixture curves for the basic stimuli of the I.C.I. system.

[7] Edward C. Dench took part in this development and is a co-author of the following paper which describes in greater detail the means devised for achieving color correction in both three- and four-color printing.

[8] A. C. Hardy and F. L. Wurzburg, Jr., "The theory of three-color reproduction," J. Opt. Soc. Am. **27**, 227 (1937).

TABLE II.

Color	Symbol	Fractional area
White	w	$(1-c)(1-m)(1-y)(1-n)$
Cyan	c	$c(1-m)(1-y)(1-n)$
Magenta	m	$m(1-c)(1-y)(1-n)$
Yellow	y	$y(1-c)(1-m)(1-n)$
Red	my	$my(1-c)(1-n)$
Green	cy	$cy(1-m)(1-n)$
Blue	cm	$cm(1-y)(1-n)$
Black	cmy	$cmy(1-n)$
Black	n	$(1-c)(1-m)(1-y)n$
Black	cn	$c(1-m)(1-y)n$
Black	mn	$m(1-c)(1-y)n$
Black	yn	$y(1-c)(1-m)n$
Black	myn	$my(1-c)n$
Black	cyn	$cy(1-m)n$
Black	cmn	$cm(1-y)n$
Black	$cmyn$	$cmyn$

would suffice in subtractive methods if idealized dyes or inks could be used. Earlier in that year, we had started the construction of a three-platen scanning machine to determine the adequacy of this relatively simple type of color correction for letter-press printing. After making a number of color reproductions, it was concluded that color corrections of a higher order are necessary, even when using the best of available inks. We therefore undertook to solve the Neugebauer equations, which hold for any set of inks; but it was not until just prior to the war that the equation-solving networks were incorporated with the scanning machine. Tests showed that these networks performed with sufficient rapidity and accuracy, and this was borne out by making three-color reproductions with full color correction. For reasons discussed below, a fourth computing channel was then added to adapt the machine to the so-called four-color processes.

FOUR-COLOR PRINTING

One of the inherent disadvantages of three-color printing is the loss of detail that ensues when the three impressions are out of register. In four-color printing, the additional black impression carries much of the detail and makes the registration less critical. For this and other reasons, relatively little three-color printing is being done today.

An extension of the reasoning that led to the development of Eqs. (1) shows that an elemental area of a four-color reproduction contains sixteen components, just double the number of a three-color reproduction. These components are listed in Table II, wherein the dot ratio of the fourth structured image is represented by n. It will be noticed that, since nine of the sixteen components are listed as black, their fractional areas may be added together on the assumption that the tristimulus values of all black areas are substantially the same. When this is done, the fraction of each elemental area of the reproduction occupied by black becomes $cmy(1-n)+n$. The Neugebauer equations for a typical four-color process then become

$$\begin{aligned} R=R'=1.000&(1-c)(1-m)(1-y)(1-n)\\ &+0.146c(1-m)(1-y)(1-n)\\ &+0.442m(1-c)(1-y)(1-n)\\ &+0.857y(1-c)(1-m)(1-n)\\ &+0.398my(1-c)(1-n)\\ &+0.064cy(1-m)(1-n)\\ &+0.033cm(1-y)(1-n)\\ &+0.023cmy(1-n)+0.023n, \end{aligned} \tag{3a}$$

$$\begin{aligned} G=G'=1.000&(1-c)(1-m)(1-y)(1-n)\\ &+0.197c(1-m)(1-y)(1-n)\\ &+0.219m(1-c)(1-y)(1-n)\\ &+0.980y(1-c)(1-m)(1-n)\\ &+0.206my(1-c)(1-n)\\ &+0.196cy(1-m)(1-n)\\ &+0.014cm(1-y)(1-n)\\ &+0.019cmy(1-n)+0.019n, \end{aligned} \tag{3b}$$

$$\begin{aligned} B=B'=1.000&(1-c)(1-m)(1-y)(1-n)\\ &+0.613c(1-m)(1-y)(1-n)\\ &+0.185m(1-c)(1-y)(1-n)\\ &+0.148y(1-c)(1-m)(1-n)\\ &+0.012my(1-c)(1-n)\\ &+0.103cy(1-m)(1-n)\\ &+0.120cm(1-y)(1-n)\\ &+0.016cmy(1-n)+0.016n. \end{aligned} \tag{3c}$$

The solution of these equations presents a new problem because the three equations contain four unknown quantities, c, m, y, and n. If n is assumed to be zero, the equations take the form of Eqs. (2) for the three-color case. At the opposite extreme, the maximum value of n is determined by the fact that c, m, and y cannot be negative. From the standpoint of good printing practice, the maximum amount of black ink should be used; and it was found that this can be accomplished in the following manner.

Let the equation-solving networks described previously with reference to Eqs. (2) be modified so as to compute values of R', G', and B' in accordance with Eqs. (3) when supplied with input signals c, m, y, and n. Let the control of c, m, and y be the same as before. The networks will then automatically solve for values of c, m, and y which satisfy Eqs. (3), provided the input signal n does not require c, m, or y to be less than zero. Now let the grid of another vacuum tube be controlled by whichever of the signals c, m, or y has the smallest value at any instant, and let the plate current of this tube supply the input signal n to the equation-solving networks. The n signal will then be just large enough to reduce at least one of the input signals c, m, or y to zero. This n signal is then used to control the dot-recording circuit during a fourth scanning operation to produce the structured image from which the black plate is made.

Four-color reproductions made by this method are, from the standpoint of accuracy in the rendition of color values, virtually indistinguishable from a three-color reproduction made under comparable conditions. However, the progressive proofs are very unlike those to which the art is accustomed, as might perhaps be expected. In the first place, if a gray scale is included with the subject, it is not reproduced at all by the color plates but only by the black plate. On the other hand, a color chart containing the eight component colors is not reproduced by the black plate, except for the black area, which is not reproduced by the color plates. The appearance of the progressive proofs of a multicolored subject is, in many ways, even more striking because of their lack of resemblance to those produced by conventional procedures. Most striking of all is the appearance of a progressive proof that has received the three-color impressions but not the black impression. The radical difference between it and a three color print of the same subject amply justifies the view that three-color and four-color processes are essentially different methods of color reproduction.

THE TWO TYPES OF COLOR CORRECTION

It is of some theoretical interest to analyze and classify the two types of color correction that are intermingled in Eqs. (1) which, for this purpose, are more conveniently rewritten in the following form:

$$\begin{aligned} R' = {} & R_w-(R_w-R_c)c-(R_w-R_m)m \\ & -(R_w-R_y)y+(R_w-R_m-R_y+R_{my})my \\ & +(R_w-R_c-R_y+R_{cy})cy \\ & +(R_w-R_c-R_m+R_{cm})cm \\ & -(R_w-R_c-R_m-R_y+R_{my} \\ & \qquad +R_{cy}+R_{cm}-R_{cmy})cmy, \end{aligned} \tag{4a}$$

$$\begin{aligned} G' = {} & G_w-(G_w-G_c)c-(G_w-G_m)m \\ & -(G_w-G_y)y+(G_w-G_m-G_y+G_{my})my \\ & +(G_w-G_c-G_y+G_{cy})cy \\ & +(G_w-G_c-G_m+G_{cm})cm \\ & -(G_w-G_c-G_m-G_y+G_{my} \\ & \qquad +G_{cy}+G_{cm}-G_{cmy})cmy, \end{aligned} \tag{4b}$$

$$\begin{aligned} B' = {} & B_w-(B_w-B_c)c-(B_w-B_m)m \\ & -(B_w-B_y)y+(B_w-B_m-B_y+B_{my})my \\ & +(B_w-B_c-B_y+B_{cy})cy \\ & +(B_w-B_c-B_m+B_{cm})cm \\ & -(B_w-B_c-B_m-B_y+B_{my} \\ & \qquad +B_{cy}+B_{cm}-B_{cmy})cmy. \end{aligned} \tag{4c}$$

Color Correction of the First Type

The nature of the first type of color correction will be obvious when it is recognized that, even when the separation negatives properly evaluate the colors of the subject, they inevitably do so in terms of a set of primaries which are not the reproduction primaries of the system. In our earlier paper,[8] we outlined an experimental procedure for identifying the reproduction primaries in any subtractive process, the mathematical equivalent of which is as follows. In some area of the reproduction, let the dot ratio c be reduced by an amount Δc without altering the values of m or y. This reduction in the size of the cyan dots will increase R' by an amount $\Delta R'$, G' by an amount $\Delta G'$, and B' by an amount $\Delta B'$, where

$$\begin{aligned} \Delta R' &= -(\partial R'/\partial c)\Delta c, \\ \Delta G' &= -(\partial G'/\partial c)\Delta c, \\ \Delta B' &= -(\partial B'/\partial c)\Delta c. \end{aligned}$$

Hence, the tristimulus values of the reproduction primary controlled by the size of the cyan dot are represented by the negative values of the partial derivatives of Eqs. (4) with respect to c. Similarly, the tristimulus values of the primaries controlled by the magenta dot and by the yellow dot can be found by partial differentiation of

these equations with respect to m and with respect to y.

Assume for the moment that the printing process uses ideal inks whose spectrophotometric characteristics show no gaps between their absorption bands and no overlapping of their absorption bands. It is easily shown that, for such ideal inks, the coefficients of my, cy, cm, and cmy in Eqs. (4) are all identically zero. Hence, in this case, perfect reproduction will result when

$$R=R'=R_w-(R_w-R_c)c-(R_w-R_m)m-(R_w-R_y)y, \quad (5a)$$

$$G=G'=G_w-(G_w-G_c)c-(G_w-G_m)m-(G_w-G_y)y, \quad (5b)$$

$$B=B'=B_w-(B_w-B_c)c-(B_w-B_m)m-(B_w-B_y)y. \quad (5c)$$

For these ideal inks, the tristimulus values of the reproduction primary controlled by the cyan dot are seen by partial differentiation to be

$$-(\partial R'/\partial c)=(R_w-R_c),$$
$$-(\partial G'/\partial c)=(G_w-G_c),$$
$$-(\partial B'/\partial c)=(B_w-B_c).$$

Similarly, the tristimulus values of the other primaries are (R_w-R_m), (G_w-G_m), (B_w-B_m) and (R_w-R_y), (G_w-G_y), (B_w-B_y), respectively. In other words, the reproduction primaries of this system are represented by the coefficients of the c, m, and y terms in Eqs. (5).

Now, knowing what the reproduction primaries are, it is theoretically possible to simplify matters by imposing a further restriction on the separation negatives. Instead of requiring merely that the separation negatives properly evaluate the colors of the subject in terms of any arbitrary set of basic stimuli, let the choice of filter-emulsion combinations be such as to evaluate the colors of the subject directly in terms of the amounts of the three reproduction primaries. This could be effected by selecting filter-emulsion combinations whose spectral sensitivity curves correspond to the color-mixture curves that a normal observer would obtain if he were to use the reproduction primaries as his basic stimuli. Such separation negatives would require no color correction of the first type; and it would be quite proper to make the cyan plate from the red negative, the magenta plate from the green negative, and the yellow plate from the blue negative. Of course, if the reproduction primaries are real colors, the spectral sensitivities of the corresponding filter-emulsion combinations must be negative at certain wave-lengths.

Equations (5) indicate how separation negatives, whose spectral sensitivities are everywhere positive, can be made to serve in the production of a faithful result with a set of real reproduction primaries. By making Eqs. (5) explicit for c, m, and y, it is found that c, for example, is not a function of R alone but is, in fact, a function of both G and B as well. Those who are accustomed to thinking in colorimetric terms will recognize that these equations represent the well-known linear transformation by which data expressed in terms of one colorimetric system are translated into corresponding data in another system. It is evident, therefore, that the second, third, and fourth terms in Eqs. (4) effect color corrections of the first type which, as our earlier paper indicated, must be made even in additive processes.

Color Correction of the Second Type

Our earlier paper introduced the concept of *stable primaries* to describe the result of using ideal dyes or inks in a subtractive process. With reference to the present analysis, partial differentiation of Eqs. (4) indicates just how unstable the primaries are when physically realizable inks are used. Thus, the primary controlled by the cyan dot has tristimulus values as follows:

$$-(\partial R'/\partial c)=(R_w-R_c)-(R_w-R_c-R_y+R_{cy})y-(R_w-R_c-R_m+R_{cm})m+(R_w-R_c-R_m-R_y+R_{my}+R_{cy}+R_{cm}-R_{cmy})my, \quad (6a)$$

$$-(\partial G'/\partial c)=(G_w-G_c)-(G_w-G_c-G_y+G_{cy})y-(G_w-G_c-G_m+G_{cm})m+(G_w-G_c-G_m-G_y+G_{my}+G_{cy}+G_{cm}-G_{cmy})my, \quad (6b)$$

$$-(\partial B'/\partial c)=(B_w-B_c)-(B_w-B_c-B_y+B_{cy})y-(B_w-B_c-B_m+B_{cm})m+(B_w-B_c-B_m-B_y+B_{my}+B_{cy}+B_{cm}-B_{cmy})my. \quad (6c)$$

As was indicated above, the coefficients of the y, m, and my terms are found to be zero when ideal inks are assumed, and the tristimulus values of this primary are therefore constant. When non-ideal inks are used, the coefficients of these terms make allowance for the overlapping of the absorption bands of the three inks. It should be noted that the two types of color correction are completely intermingled in Eqs. (1), which, because of their symmetry, are better adapted for solution by electronic means.

COLOR CORRECTION IN FOUR-COLOR PROCESSES

It has been seen that, as a practical expedient, a four-color process can often be regarded as having nine components instead of sixteen. Indeed, if the black ink is black enough, the coefficients of all the terms involving n as a multiplier can be assumed to be zero. Let this assumption be made in the interest of gaining some insight into the nature of the color corrections in this type of process.

Inspection of Eqs. (3) will show that, with this assumption, the equations are exactly like those of the three-color case (Eqs. (2)), except for the common factor $(1-n)$. Hence, the right-hand sides of Eqs. (3) become identical with those of Eqs. (2) when multiplied by $1/(1-n)$. Now, consider the physical significance of multiplying R, G, and B by this quantity $1/(1-n)$, whose numerical value is always equal to or greater than unity. Imagine a subject that is to be reproduced by a four-color process, and imagine another subject like the first, except that the tristimulus values of every elemental area of the second are $1/(1-n)$ times those of the first. This modified subject will be more luminous than the original, but everywhere of the same chromaticity. Indeed, on the assumption that $cmy=0$, the luminance of every elemental area will be increased until it is equal to the maximum luminance attainable at this chromaticity by the given set of colored inks, used in pairs. This modified subject represents the appearance that the four-color reproduction should have before the final black impression is made.

The literature pertaining to the art of four-color printing contains many erroneous statements concerning the amount of dot-size reduction required in the color plates of a four-color set. The straightforward method of determining the amount of reduction required would be to solve the Neugebauer equations for values of c, m, and y in the three-color case and then to solve them again in the four-color case, making use of the additional fact that $cmy=0$. A comparison of the results would then indicate the amount of reduction required in any area of the subject whose tristimulus values are R, G, and B. The fact that these algebraic operations are too formidable to be considered seriously need not deter us from the conclusion that there are no simple rules by which a set of three-color plates can be converted for use in a four-color process.

MULTICOLOR PROCESSES

The electronic methods that have been developed for solving the Neugebauer equations have proved so satisfactory that extension to processes using more than four colors should involve no real difficulties. Of course, there will always be only three fundamental equations, but they can be supplemented by the imposition of additional conditions, as was done in the treatment of the four-color case.

It is of interest in this connection to note again that, regardless of the number of impressions, the inks may be selected solely on the basis of their color gamut. Their colors need not be cyan, magenta, and yellow; nor is it required that they be transparent. The way is therefore opened for entirely new printing processes.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 38, NUMBER 4 APRIL, 1948

An Electronic Method for Solving Simultaneous Equations[1]

ARTHUR C. HARDY
Massachusetts Institute of Technology, Cambridge, Massachusetts

AND

EDWARD C. DENCH[2]
Interchemical Corporation, New York, New York

(Received December 22, 1947)

In connection with the requirements of color correction discussed in the preceding paper, we found that a set of n simultaneous equations can be solved by using each of the n error signals to control one of the independent variables. Stability of this equation-solving network can be achieved by the proper interconnection. In applying this procedure to the equations for color correction, it was necessary to find the sum of eight terms, each of which is the product of either three or four independent variables. It was found that these individual products could be computed by making use of the fact that, if n events have probabilities of occurrence a, b, $\cdots n$, respectively, the probability of their simultaneous occurrence is the product of their individual probabilities $abc\cdots n$. By generating rectangular waves of irrationally related frequencies whose positive pulse widths are proportional to the values of the unknown variables, and applying these signals to the grid of a vacuum tube in such a manner that its plate current flows only when all of the signals have positive values simultaneously, the average value of the plate current is proportional to the required product. When controlled in this way, eight vacuum tubes with their plates connected in parallel deliver an average current that is proportional to the required sum.

THE preceding paper[3] has shown that it is possible to base an automatic method of color correction on a set of three simultaneous equations of either third degree or fourth degree. The purpose of this paper is to describe the electronic methods employed in solving these equations continuously at a rate that is satisfactory for recording the dots on structured images by scanning,—a rate that corresponds to approximately a thousand solutions per second.

SOLUTION OF THE THREE-COLOR EQUATIONS

In a typical case, the three-color equations to be solved are:

$$\begin{aligned}R=R'=&1.000(1-c)(1-m)(1-y)\\&+0.146c(1-m)(1-y)\\&+0.442m(1-c)(1-y)\\&+0.857y(1-c)(1-m)\\&+0.398my(1-c)+0.064cy(1-m)\\&+0.033cm(1-y)+0.023cmy,\end{aligned}\quad(1a)$$

$$\begin{aligned}G=G'=&1.000(1-c)(1-m)(1-y)\\&+0.197c(1-m)(1-y)\\&+0.219m(1-c)(1-y)\\&+0.980y(1-c)(1-m)\\&+0.206my(1-c)+0.196cy(1-m)\\&+0.014cm(1-y)+0.019cmy,\end{aligned}\quad(1b)$$

$$\begin{aligned}B=B'=&1.000(1-c)(1-m)(1-y)\\&+0.613c(1-m)(1-y)\\&+0.185m(1-c)(1-y)\\&+0.148y(1-c)(1-m)\\&+0.012my(1-c)+0.103cy(1-m)\\&+0.120cm(1-y)+0.016cmy.\end{aligned}\quad(1c)$$

From the standpoint of this paper, R, G, and B can be assumed to represent three photo-tube signals which vary continuously during the scanning operation, in correspondence with color variations from point to point of the original. Given a set of constants (which depend upon the inks used and the printing conditions), the problem is to find values of c, m, and y which satisfy these equations for any set of realizable values of R, G, and B. As was pointed out in the previous paper, the solutions for c, m, and y are to be in the form of electrical signals that can be used to control the dot-producing circuit.

It is difficult to describe the behavior of circuits that solve simultaneous equations continuously because of the number of operations which, of necessity, occur simultaneously. A simplified

[1] Extension of paper presented at the October, 1946 meeting of the O.S.A.
[2] Now with Raytheon Manufacturing Company.
[3] A. C. Hardy and F. L. Wurzburg, Jr., J. Opt. Soc. Am. 16, 300 (1948).

version of the procedure that we have devised will be described with reference to Fig. 1, and it should be noted that this procedure is generally applicable to the solution of any set of simultaneous equations. The plate currents of the three vacuum tubes shown in this figure will be designated by c, m, and y because, as will be evident later, they ultimately represent these quantities. These plate currents are fed into three computing channels (to be described later), where they are combined in such a manner that the outputs of the computing channels are signals representing R', G', and B', as defined by Eqs. (1). The set of simultaneous equations is solved when each of these signals is equal to the corresponding photo-tube signal, R, G, or B.

Consider now the computing channel whose output signal is R'. This signal is compared with the photo-tube signal R by means of a subtractor unit, the output of which is proportional to the difference between R and R'. This difference or error signal is fed into a high gain amplifier A, whose output controls the current c in such a manner that the value of R' tends to satisfy the condition $R'=R$ of Eq. (1a). It will be seen from Fig. 1 that the currents m and y are controlled in a similar manner by the error signals $(G-G')$ and $(B-B')$, respectively. Thus, m and y tend to assume values which satisfy the conditions $G'=G$ and $B'=B$ of Eqs. (1b) and (1c). If there is sufficient gain in the high gain amplifiers, the currents c, m, and y take on values that represent the desired solutions of the equations.[4]

THE COMPUTING CHANNELS

It was assumed above that the output signals of the three computing channels represent R', G', and B', respectively, when they are supplied with input signals c, m, and y. In other words, each computing channel was assumed to perform the operations indicated by the equation to which it corresponds.

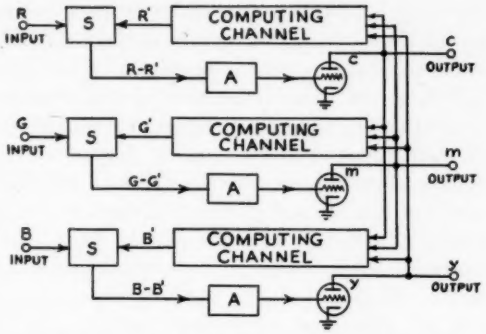


FIG. 1. Schematic circuit diagram illustrating method for solving simultaneous equations.

Consider the problem of computing R' in accordance with the operations indicated by Eq. (1a). This equation contains eight terms to be added together, each term being a different function of c, m, and y.[5] There are eight vacuum tubes in the computing channel for R', each corresponding to one of the terms in the equation. The plate currents of these tubes flow through a common load resistor and, at any given instant, one (and only one) of the eight tubes delivers plate current, the others being biased below cut-off. The average current flowing through this common load resistor can be made to represent R' if the magnitude of the plate current of each tube (while it is conducting) represents the coefficient of the corresponding term in Eq. (1a) and if each tube is allowed to deliver plate current intermittently for an appropriate fraction of the time. For example, the last term of Eq. (1a) is $0.023cmy$. The coefficient of this term can be represented by adjusting the bias of the screen grid of the corresponding tube until, with its control grid at the potential of the cathode, the plate current is 0.023 (arbitrary units). If the input signals to the computing channels happen to be $c=m=y=\frac{1}{2}$, the numerical value of the product cmy is $\frac{1}{8}$. Thus, for these input signals, the numerical value of the last term of Eq. (1a) is

[4] It should be mentioned that this mode of interconnection does not lead to a solution of the equations for all possible values of the constants. However, it can be surmised from the nature of the procedure in color printing that R' will generally depend to a greater extent on the value of c than on the values of m and y. Also, since G' is chiefly dependent on m and B' is chiefly dependent on y, the mode of interconnection described is, in fact, the preferred mode for the purpose at hand. Interest in the stability of multiple servo-networks was so stimulated by the war that it seems not unlikely that it may someday be possible to select the preferred mode of interconnection for any set of equations without reference to the underlying physical process or without trying the several possible modes of interconnection experimentally.

[5] It was shown in the preceding paper that c, m, and y represent dot ratios which always have fractional values lying between 0 and 1. The numerical values of $(1-c)\times(1-m)(1-y)$, $c(1-m)(1-y)$, etc., are always positive fractions whose sum is unity.

$0.023 \times \frac{1}{8}$, and this term can be properly represented if the corresponding tube delivers its 0.023 units of plate current intermittently for one-eighth of the time. Inspection of the equation shows that, for this particular set of input signals, all eight tubes in the computing channel should deliver plate current for one-eighth of the time also; but this is a special case that does not occur with any other set of input signals. Since what has been said here with reference to the computation of R' applies also to the computation of G' and B', the three computing channels can be identical in construction.

The eight functions of c, m, and y that appear in these equations are all triple products; and the multiplication of electrical signals to obtain even a single product has hitherto required an amount of apparatus that would have been prohibitive, at least in this sort of application. We therefore undertook to find a simpler method of multiplication, preferably one based on some fundamental law rather than on the characteristics of thermionic tubes under special conditions of operation. It is known from the theory of probability that, if three independent events have probabilities of occurrence p, q, and r, the probability of their simultaneous occurrence is pqr. It therefore occurred to us that, if each of the three vacuum tubes of Fig. 1 were replaced by a signal generator of the on-off type, such as a multivibrator or a constant-amplitude pulse generator, the time duration of the signal pulses during each cycle could easily be made to represent c, m, and y, respectively. Then, if the generator frequencies are irrationally related, the generation of a pulse by any one of them can be regarded as an independent event. By applying the three signals to the grid circuit of a tube in the computing channel in the manner illustrated in Fig. 2, the tube can easily be kept biased below cut-off except during periods when positive signal pulses are being received simultaneously from the three generators. When so controlled, the fraction of the time during which this tube delivers plate current is cmy. This solves the problem of multiplication indicated by the last term in each of the three equations.

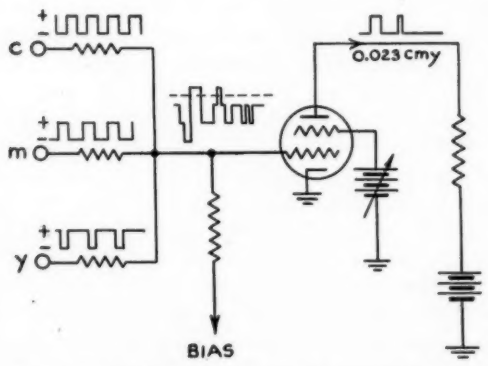


FIG. 2. Schematic circuit diagram illustrating a method, based upon the theory of probability, for obtaining the product of three (or more) electrical signals. The input signals c, m, and y are supplied by pulse generators operating at irrationally related frequencies, the numerical values of the quantities being represented by the fractional lengths of the positive pulses. These signals are added and applied to the grid of a vacuum tube in the manner illustrated, the grid being biased below cut-off except when the three input signals are positive simultaneously. The plate current is intermittent but its average value is proportional to the product cmy. This diagram illustrates the computation of the last term in Eq. (1a), the proportionality constant, 0.023, being introduced by proper adjustment of the bias of the screen grid.

The control of the other tubes in the computing channels is effected by using an inverter stage in company with each pulse generator to produce positive signal pulses whose time durations represent $(1-c)$, $(1-m)$, and $(1-y)$, respectively. These signals, like the signals representing c, m, and y, are carried on "bus bars" to which the grid circuits of the tubes in the computing channels are connected in the required manner. That is, the grid circuit of the first tube in each computing channel is connected to the bus bars carrying the $(1-c)$, $(1-m)$, and $(1-y)$ signals, the second tube in each channel is connected to the c, $(1-m)$, and $(1-y)$ bus bars, etc. Because of the simplicity of this arrangement, the three computing channels occupy about as much space as the average portable radio set.

ACCURACY OF THE METHOD

At the time the above circuits were constructed, the choice of frequencies for the pulse generators represented a compromise between speed and accuracy. On the one hand, it is desirable that the generator frequencies be high by comparison with the rate at which dots are to be produced. On the other hand, prewar techniques did not enable this type of signal to be generated at high

frequencies without rounding the corners of the "square waves;" and this rounding would introduce errors in the computation. As a compromise, we employed irrationally related signal frequencies in the neighborhood of 18,000, 21,000, and 24,000 cycles per second, and were pleased to discover that the errors in computation were significantly less than other errors that occur in color printing, even when dots were recorded at a rate of more than one thousand per second. If it were necessary, electronic techniques developed during the war would no doubt enable both the accuracy and the speed of operation to be considerably increased.

It should be mentioned that linearity in the high gain amplifiers is not essential, and that the gain in these amplifiers can be varied over rather wide limits without appreciably altering the solutions of the equations. It should be mentioned also that the constants in the equations can be determined on the scanning machine itself. A color chart containing the eight colors of the three-color process is photographed with the original; and the potentiometers in the computing channels are adjusted, with the scanning beams on the images of appropriate areas of the color chart, until the signals representing c, m, and y have their known values. That is, white represents $c=m=y=0$, cyan represents $c=1$, $m=0$, $y=0$, etc. When the circuits have been so adjusted that these eight colors will be reproduced correctly, all other colors within the attainable gamut can be reproduced correctly by any printing process to which the Neugebauer equations apply.

SOLUTION OF THE FOUR-COLOR EQUATIONS

In a typical case, the four-color equations are:

$$\begin{aligned} R=R'=1.000&(1-c)(1-m)(1-y)(1-n)\\ &+0.146c(1-m)(1-y)(1-n)\\ &+0.442m(1-c)(1-y)(1-n)\\ &+0.857y(1-c)(1-m)(1-n)\\ &+0.398my(1-c)(1-n)\\ &+0.064cy(1-m)(1-n)\\ &+0.033cm(1-y)(1-n)\\ &+0.023cmy(1-n)+0.023n, \end{aligned} \quad (2a)$$

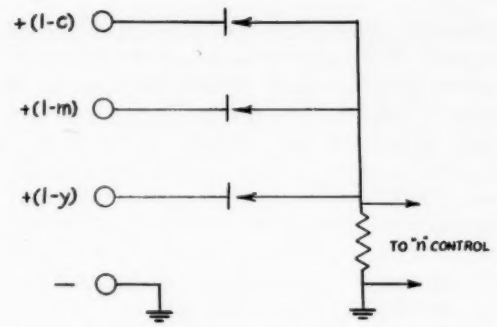


FIG. 3. Schematic diagram of maximum selector used to control the fourth (black) signal generator.

$$\begin{aligned} G=G'=1.000&(1-c)(1-m)(1-y)(1-n)\\ &+0.197c(1-m)(1-y)(1-n)\\ &+0.219m(1-c)(1-y)(1-n)\\ &+0.980y(1-c)(1-m)(1-n)\\ &+0.206my(1-c)(1-n)\\ &+0.196cy(1-m)(1-n)\\ &+0.014cm(1-y)(1-n)\\ &+0.019cmy(1-n)+0.019n, \end{aligned} \quad (2b)$$

$$\begin{aligned} B=B'=1.000&(1-c)(1-m)(1-y)(1-n)\\ &+0.613c(1-m)(1-y)(1-n)\\ &+0.185m(1-c)(1-y)(1-n)\\ &+0.148y(1-c)(1-m)(1-n)\\ &+0.012my(1-c)(1-n)\\ &+0.103cy(1-m)(1-n)\\ &+0.120cm(1-y)(1-n)\\ &+0.016cmy(1-n)+0.016n, \end{aligned} \quad (2c)$$

where all quantities have the same meaning as before and n is the dot ratio of the black dot. Because these equations contain four unknown quantities, they can be satisfied by an infinite set of values for c, m, y, and n. Imposing the condition that $n=0$ reduces the equations to the form of those of the three-color case. On the other hand, imposing the condition that $cmy=0$ requires that at least one of the color dots be of zero size in every area of the reproduction. Our experiments with four-color printing have assumed the latter condition, except that, as a practical matter, the dots of a halftone plate must never be allowed to drop out entirely.

Our four-color circuits are identical with the three-color circuits, except for the addition of another color channel and its control circuit. This color channel supplies signals representing n and $(1-n)$, which are carried on bus bars into the

computing channels, each tube of the computing channels being so biased that it delivers plate current only when four positive signal pulses are being received simultaneously. The control of the color channel whose output represents n is effected by means of the circuit shown schematically in Fig. 3. The $(1-c)$, $(1-m)$, and $(1-y)$ signals, after being filtered, pass through rectifier units and then through a common resistor to ground. Because of the presence of the rectifiers, the signal voltage developed across the resistor is determined by whichever of the input signals is greatest. If, for example, $(1-c)$ is greater than either $(1-m)$ or $(1-y)$, c must obviously be less than either m or y. The signal voltage developed across the resistor is amplified by another high gain amplifier and is used to control the pulse generator producing the n signal. The effect is to introduce as much black into the reproduction as is consistent with the requirement that c, m, and y must never be less than zero or some predetermined minimum.

ACKNOWLEDGMENT

The authors are pleased to acknowledge the able assistance of C. B. Kinne and H. E. Rose in the construction and testing of the electronic equipment.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 38, NUMBER 4 APRIL, 1948

Absolute Sensitivity Measurements on Single-Grain-Layer Photographic Plates for Different Wave-Lengths *

J. H. WEBB
Kodak Research Laboratories, Rochester 4, New York
(Received October 10, 1947)

Single-grain-layer photographic plates were prepared from a non-color-sensitized, slow-speed, uniform-grain-size emulsion and exposed to very pure monochromatic light of wave-lengths λ3600A, λ4500A, and λ5000A. Intensity-scale characteristic curves were made on a double monochromator in which the energy of the exposures could be measured with an accuracy of ±5 percent. Data for plotting the characteristic curves were obtained by microscopic grain counts of developed and undeveloped grains for successive exposure steps. These were then plotted in several different ways: First, the fractions of developed grains, k/N (k is the developed and N the total number of grains per unit area) were plotted as a function of the incident exposure in terms of ergs per square centimeter of plate area. Second, the k/N values were plotted as a function of incident exposure in terms of quanta per average grain. Finally, using absorption data obtained for thin layers of silver bromide, the k/N values were plotted as a function of quanta absorbed per grain. Experimentally, it is found that the characteristic curves increase their slope with increasing wave-length of exposing radiation and also show a tendency to increase in slope with decreasing development time.

The characteristic curves are analyzed by means of the Poisson statistical equation to determine the number of quanta actually utilized in formation of the latent image. Preparatory to this analysis, exposures were made to alpha-particles, and it is shown that the single-grain-layer plates accurately obey the single-hit formula derived from the Poisson equation. From the analysis of the curves obtained with monochromatic light exposures, it is concluded that the threshold sensitivity of a grain, expressed as the minimum number of quanta to produce developability of the grain, is of the order of ten quanta per grain for the emulsion under test. The upper limit of sensitivity is specified by the number of quanta absorbed per grain as shown by the abscissa values of the characteristic curves.

Reciprocity-law failure measurements were made on the single-grain-layer plates, using unfiltered tungsten radiation. Time-scale characteristic curves were made over an intensity range of 1:16,384. The curves showed some loss of speed with reduced intensity but mainly a reduction in slope. This indicates that the grains of the upper and lower parts of the characteristic curve do not behave the same toward reduced intensity, and the grains of the upper part of the curve show the greater loss in speed at low intensity.

INTRODUCTION

A GREAT deal of work has been done in photographic latent-image theory using ordinary multigrain-layer emulsions as the experimental medium. In such experiments, it is customary to draw conclusions about the effects of different chemical agents and exposing procedures from the densities measured by the

usual methods on the developed multigrain-layer emulsions. It is appreciated by workers in latent-image theory that the practice of using thick emulsions for many types of experiments is far from desirable because of the complicating conditions thereby imposed on the processes of exposure and development, and in the interpretation of results. If the purpose of the experiment in photographic theory is to study what happens in individual photographic grains upon exposure under controlled conditions, it is obviously desirable to have the grains to be studied spread out in a single layer and separated sufficiently so that they do not interfere with each other during exposure and development, and during the subsequent grain-counting procedure.

A great deterrent to the use of single-grain-layer plates in latent-image work is the tedious nature of the work of preparation of the plates and of evaluating the results of exposure. Once prepared, the single-grain-layer plates are exposed and developed in a more or less conventional manner, but the evaluation of the effects of exposure must be carried out by microscopic grain counts of developed and undeveloped grains. To obtain reliable data, large numbers of grains (several hundred) must be counted for each exposure step. When it is considered that five to ten separate exposure steps are required to obtain one characteristic curve, it is apparent that any extended investigation involving many characteristic curves will become a great task.

Notwithstanding the laborious work involved, a number of workers have carried out photographic investigations using the single-grain-layer technique. Among the more important of these should be mentioned the work of Scheffer[1] on photomicroscopy of developing and developed grains; the work of Hodgson,[2] who studied partial development of photographic grains; of Svedberg,[3] who studied partial development of grains, effects of hydrogen peroxide on the photographic grains, and distribution of sensitivity centers on the surfaces of grains; the work of Toy[4] on variation of sensitivity among grains and distribution of sensitivity centers on the surfaces of grains; the work of Sheppard, Trivelli, and Loveland[5] on photographic sensitivity and formation of the latent image; of Wightman, Trivelli, and Sheppard[6] on the effect of hydrogen peroxide on single photographic grains; the work of Silberstein and Trivelli[7] on the quantum theory of x-ray exposures; the work of Trivelli and Sheppard[8] on visible decomposition of the silver halide grains by light; the work of Trivelli and Righter[9] on the quantum theory of photographic exposure; the work done by Trivelli and Loveland[10] on the effect of grain size on reciprocity-law failure; and the extensive theoretical investigations by Silberstein[11] on the quantum theory of exposure.

Certain types of experiments almost demand the use of single-grain-layer plates if indisputable results are to be obtained. The main experiment to be reported here, consisting of the determination of the sensitivity of individual photographic grains at different wave-lengths, is considered to be in this category. Accordingly, in the present work a limited experimental program was undertaken using this technique. The emulsion used was of very uniform grain size, the exposures were made with monochromatic light of high purity, and the energy of the exposing radiation was measured with an accuracy of the order of plus or minus 5 percent. Exposures were made at wave-lengths λ3600A, λ4500A, and λ5000A, which covered the complete sensitivity range of the unsensitized emulsion used.

EXPERIMENTAL PROCEDURE

Single-Grain-Layer Plates

The emulsion used was a non-color-sensitized, low-speed, fine-grain, positive-type emulsion,

* Communication No. 1158 from the Kodak Research Laboratories.

[1] W. Scheffer, Brit. J. Phot. **54**, 116, 271 (1907).

[2] M. B. Hodgson, J. Frank. Inst. **184**, 705 (1917).

[3] T. Svedberg, Zeits. wiss. Phot. **20**, 36 (1920); *ibid.*, Phot. J. **62**, 186 (1922).

[4] F. C. Toy, Phot. J. **61**, 417 (1921); *ibid.*, Phil. Mag. **44**, 365 (1922); *ibid.*, Phil. Mag. **45**, 715 (1923).

[5] S. E. Sheppard, A. P. H. Trivelli, and R. P. Loveland, J. Frank. Inst. **200**, 51 (1925).

[6] E. P. Wightman, A. P. H. Trivelli, and S. E. Sheppard, J. Frank. Inst. **200**, 335 (1925).

[7] L. Silberstein and A. P. H. Trivelli, Phil. Mag. **9**, 787 (1930).

[8] A. P. H. Trivelli and S. E. Sheppard, J. Phys. Chem. **29**, 1568 (1925).

[9] A. P. H. Trivelli and L. Righter, Phil. Mag. **44**, 252 (1922).

[10] A. P. H. Trivelli and R. P. Loveland, J. Frank. Inst. **209**, 639 (1930).

[11] L. Silberstein, Phil. Mag. **44**, 257 (1922); *ibid.*, Phil. Mag. **45**, 1062 (1923); *ibid.*, Phil. Mag. (Series 7) **5**, 464 (1928).

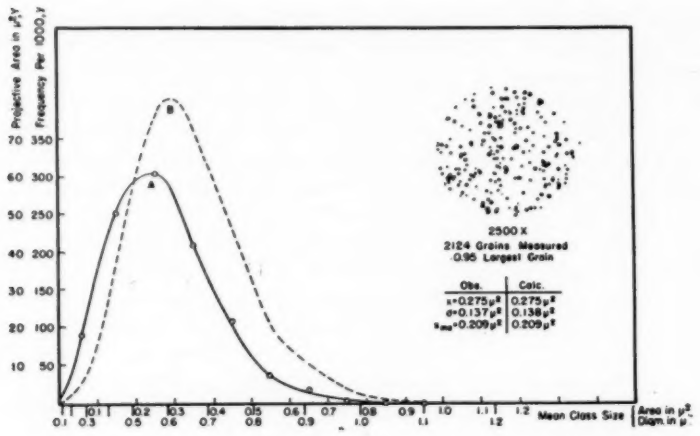


FIG. 1. Grain size-frequency curve for test emulsion, A = size frequency per 1000 grains, B = size total projective area per 1000 grains, $Y = Xy$, $\Delta X = 0.10\mu^2$.

having a uniform grain size. The composition of the grains is mainly silver bromide with 2 mole percent silver iodide. A photomicrograph of the emulsion, together with a grain size-frequency curve, is shown in Fig. 1. It may be seen that the grains appear uniform in size, and spherically shaped. The average area of a single grain as read from the figure is $0.275\mu^2$ (0.275×10^{-8} cm²). The single-grain layers were prepared by diluting a sample of regular production emulsion and spreading it on optically polished glass plates 2 inches by 5 inches in size. These glass plates had a red dye antihalation backing to prevent exposure by back-reflected light. The grain density per square centimeter of plate area, as measured microscopically, varied between the limits of 5.7×10^6 and 23×10^6 grain per square centimeter. The average value was of the order of 15×10^6 grain per square centimeter, which corresponds to a coverage factor of approximately 0.0423 of the plate area.

Exposure

The exposures were carried out using a double monochromator. This instrument employs two replica gratings as dispersing elements and gives a narrow band of monochromatic light very free of scattered light.[12] The wave-length band transmitted through the slits at each setting of the instrument is about 50A in width. Tests for stray light have shown that when the instrument is set in the red part of the spectrum, and a film having red sensitivity only 10^{-5} times its blue sensitivity is being measured, the amount of stray blue radiation present will cause an error of about 10 percent. This is a severe test and, for the measurements carried out here, where the relative sensitivities at the wave-lengths employed cover a range of only $1:10^{-3}$, the error from stray light is less than 0.1 percent.

The exposures were made by varying the intensity factor of exposure and consisted of ten power-of-two steps made simultaneously at an exposure time of 15 seconds. The exposing light consists of a magnified image of the exit slit of the monochromator with tenfold linear magnification. A movable thermopile, which is built into the instrument, can be placed immediately behind the exit slit when an energy measurement is desired. The thermopile is a type F109 Hilger thermopile and was calibrated by using a standard radiation lamp supplied and certified by the National Bureau of Standards. A curve showing intensity *versus* wave-length is shown in Fig. 2. The curve at shorter wave-lengths was obtained with no screen in the beam, whereas the curve from λ4000A to λ7000A was taken with a wire-mesh screen placed at the condenser lens in front of the light source. The two overlapping measurements at wave-length λ4000A will serve to show the transmission of the screen. The energy curve shown represents the radiation distribution when a ribbon-filament tungsten lamp

[12] For description of instrument, see C. H. Evans, "An intensity-scale monochromatic sensitometer," J. Opt. Soc. Am. **30**, 118 (1940).

is used as source in this instrument. The intensity of the exposing radiation at the plate position, I_p, can be obtained from the galvanometer deflection by use of the constant given under Fig. 2. The accuracy of the intensity measurements of the radiation at the exit slit of the instrument by means of the thermopile is believed to be within the limits of ±5 percent.

Development

The plates were developed in D-16 developer for different specified periods of time, at a temperature of 68°F. Development on the single-grain-layer plates was stopped by immersion of the plate in a 1 percent solution of acetic acid, after which the plates were thoroughly washed and dried. Plates were not fixed, in order that the undeveloped grains would remain for counting.

A point which should be emphasized here is the low fog value of the single-grain-layer plates when freshly prepared. From a practical standpoint, there were no fog grains when the plates were freshly made. After a period of 6 or 7 weeks some fog grains began to appear, but the experiments were carried out prior to this time.

CALIBRATION OF GRAIN-COUNTING PROCEDURE BY MEANS OF ALPHA-PARTICLE EXPOSURE

It is well known that alpha-particles from ordinary natural sources (4–8 Mev) are practically 100 percent effective in producing developability of photographic grains. It was first demonstrated by Kinoshita[13] that the density on the photographic plate resulting from an exposure to N alpha-particles per square centimeter follows the law,

$$\frac{D}{D_{\max}}=(1-e^{-cN}),$$

in which c is a constant, and, further from absolute sensitivity measurements, that each photographic grain can be made developable by the action of one alpha-particle. Later, Svedberg and Anderson[14] showed that if a single layer of photographic grains of uniform size, and spread haphazardly on a surface, is subjected to a stepped series of increasing exposures to alpha-particles, the fraction of grains brought up to the developable state follows the Poisson statistical law. According to the Poisson law of fluctuation of small numbers, if a photographic grain of area a square centimeters receives on the average an quanta, the probability that it receives exactly j quanta is

$$P_{(j)}=\frac{e^{-an}(an)^j}{j!}. \tag{1}$$

Therefore, if a large number of grains, N, spread haphazardly over the unit area of a plate, is subjected to an exposure of n particles per square centimeter, the number of grains receiving j particles will be

$$k=NP_{(j)}, \tag{2}$$

and those receiving r or more particles will be

$$k=N\sum_{j=r}^{\infty}P_{(j)}=N\sum_{j=r}^{\infty}\frac{e^{-an}(an)^j}{j!}$$

$$=N\left[1-e^{-an}\sum_{j=0}^{r-1}\frac{(an)^j}{j!}\right]. \tag{3}$$

If r is assumed equal to unity, one obtains the one-particle-hit formula,

$$\frac{k}{N}=(1-e^{-an}), \tag{4}$$

which was verified by Svedberg and Anderson[14] for photographic grains exposed to alpha-particles.

In order to check on various phases of the

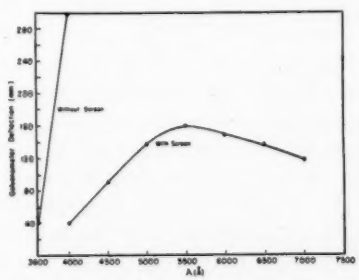


FIG. 2. Energy calibration curve for monochromator $I_p=2.44\times10^{-2}$ ergs/cm²/sec./mm deflection.

[13] S. Kinoshita, Proc. Roy. Soc. **A83**, 432 (1910).
[14] T. Svedberg and H. Anderson, Phot. J. **61**, 325 (1921).

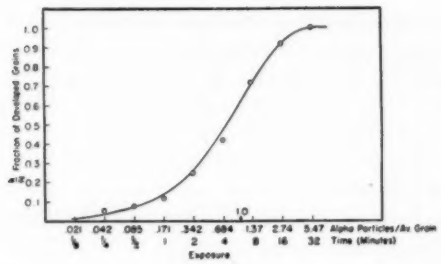


FIG. 3. Alpha-particle exposure curve for single-grain-layer plate; developed 3 min., D–16 at 68°F. ⊙ experimental, — theoretical ($r=1$).

experiment to be carried out here, some exploratory exposures were made on the test emulsion to polonium as the source of alpha-particles. The exposures covered a range of exposure times from $\frac{1}{4}$ minute to 32 minutes in eight power-of-two steps. A metal strip covered with an electrolytic deposit of polonium was placed at a distance of 1.5 cm below a second metal plate containing a small rectangular aperture for limiting the exposure area. The single-grain-layer plate was placed in contact with this upper metal plate and moved along, one step at a time, to give the different exposure values mentioned. Results of a typical exposure are shown in Fig. 3, in which the ordinates represent the fractions of grains made developable by the exposure values given along the abscissa axis. The points enclosed with circles represent the experimental values. The solid-line curve represents the theoretical curve drawn according to the Poisson equation (3), with q equal to an and r set equal to unity. This curve was selected from the group of theoretical curves shown in Fig. 4 and drawn according to Eq. (3) for values of r ranging from 1 to 100. As may be seen from Fig. 3, the experimental points of the alpha-particle exposure are very closely in agreement with the theoretical one-hit particle formula. This was, of course, found to be the case by Svedberg and Anderson.[14] The reason for repeating the experiment here was to check the microscopic grain-counting procedure for a case in which it was reasonably certain that the Poisson statistical law would be applicable. The main factor that was expected to cause divergence between experiment and theory was the grain-size variation as shown in Fig. 1. The close agreement found between experiment and theory, however, seems to indicate that the grain-size variation does not constitute a large source of error. It also served to give confidence in the technique and accuracy of the grain-counting procedure employed. These findings were considered to indicate that grain-size variation and counting technique would not introduce large errors in the experiments with light exposure to be discussed below.

Upon finding the close agreement between shapes of the experimental and theoretical curves for the alpha-particle exposure, it was decided to check the absolute sensitivity of the single-grain-layer plate when given known exposures to alpha-particles. To accomplish this it was necessary to measure the absolute intensity, number of alpha-particles per square centimeter per second, used in the alpha-particle experiment. This was done by exposing one of the Eastman NTA Plates (plate for measuring alpha-particle tracks) to the polonium source under the same conditions of exposure used on the single-grain-layer plate, and then counting the alpha-particle tracks recorded. In order to hold the number of tracks at a level such that they could readily be counted under the microscope, the exposure time was limited to 10 seconds. By the use of a 4-mm microscope objective, magnification 43×, and a micrometer eyepiece disk with a 49-mm² area for counting, an average of 31 alpha-particle tracks per square millimeter of reticle area was found. Reduced to alpha-tracks per square centimeter of plate area, and after allowing for a reduction of intensity of the polonium source because of lapse of time between experimental and calibration exposures,

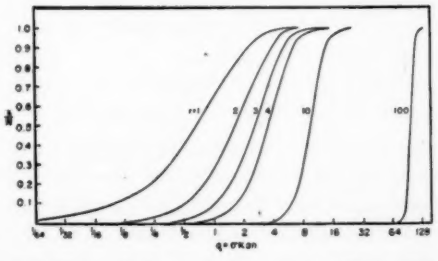


FIG. 4. Theoretical exposure curves according to Poisson equation

$$\frac{k}{N}=1-e^{-q}\sum_{j=0}^{r-1}\frac{(q)^j}{j!}.$$

this gave a value of 0.625×10^8 alpha-particles per square centimeter per minute of exposure. Absolute exposure values found in this way are shown on the abscissa axis of Fig. 3. Theoretically, the one-quantum value should occur at the point where the k/N value is 0.628. Actually, the one-quantum value falls at the k/N value of 0.6. This agreement is perhaps better than is to be expected, but it will serve to show that the single-grain-layer emulsion used in these experiments behaves statistically very well in accord with theory when exposed to a bombardment of projectiles of sufficient energy that one impact on a grain is sufficient to render that grain developable.

EXPERIMENTAL RESULTS WITH EXPOSURE TO MONOCHROMATIC LIGHT

Experimental results on the single-grain-layer plates exposed to monochromatic light of wave-lengths λ3600A, λ4500A, and λ5000A, are shown in Figs. 5, 6, and 7. In Fig. 5 the characteristic curves are portrayed as fractions of developable grains as a function of incident exposure in ergs per square centimeter of plate area. The abscissa values being given in absolute units of incident exposure, the relative positions of the curves along the exposure axis correspond to the actual sensitivity differences in the emulsion for the different wave-lengths. For each wave-length, two curves corresponding to development times of three minutes and eight minutes are shown. In addition, a curve for one and one-half minutes development time is shown for wave-length λ4500A. Referring to the three-minute development time series of curves, it may be seen that the sensitivity change between the wave-lengths

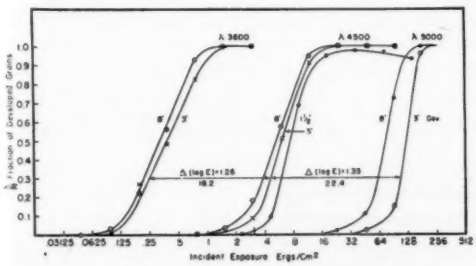


FIG. 5. Characteristic curves on single-grain-layer plates for different wave-lengths.

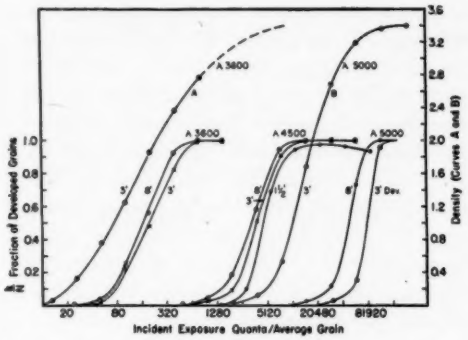


FIG. 6. Characteristic curves on single-grain-layer and multigrain-layer plates for different wave-lengths.

λ3600A and λ4500A is 1.26 in log units corresponding to a numerical factor of 18.2, and between the wave-lengths λ4500A and λ5000A is 1.35 in log units corresponding to a numerical factor of 22.4. It is also to be noted that there is some shift of the curves with varying development time as is to be expected.

The curves of Fig. 5 have been replotted in Fig. 6 as a function of incident exposure expressed as quanta per average grain. The exposure in ergs per square centimeter was converted to quanta per square centimeter by the relationship,

$$N \text{ quanta} = \frac{\lambda(A)\times E \text{ ergs}\times 10^8}{1.965}. \qquad (5)$$

The exposure in terms of quanta per average grain was obtained by multiplying the exposure per square centimeter by the average grain size $0.275\mu^2$, as given earlier.

The curves of Fig. 6 are identical in shape with those of Fig. 5, but are spread a little farther apart along the exposure axis because the relationship between quanta and energy depends upon wave-length. In addition to the single-grain-layer plate curves, two characteristic curves for wave-lengths λ3600A and λ5000A for a multilayer coating of the test emulsion are also included in this figure. These two curves show density as a function of incident exposure expressed as quanta per average grain, calculated as above. The exposures on this multilayer emulsion at wave-lengths λ3600A and λ5000A were made on the same sensitometer and under

the same conditions as were the exposures on the single-grain-layer plates. The positions of these two curves relative to the curves for the same wave-lengths on the single-grain-layer plates are to be especially noted. The curve for λ3600A on the multilayer emulsion falls very near to the curve of corresponding wave-length on the single-grain-layer plate. On the other hand, the curve for λ5000A on the multilayer emulsion falls considerably to the left of the corresponding curve on the single-grain-layer plate. This is undoubtedly connected with the fact that for the short wave-length light, the upper layers of the multilayer emulsion produce a screening action toward exposure of the lower layers of the emulsion. For the longer wave-length radiation, on the other hand, the multigrain layer emulsion being very transparent, the light penetrates readily through the emulsion. In this latter case also, the back-and-forth reflection of light between grains extends the path of the radiation manyfold in the multilayer emulsion and thus gives the light a chance to pass back and forth through individual grains many times. The net effect of this is to increase greatly the efficiency of utilization of long wave-length radiation by a multilayer emulsion over that of a single-grain-layer plate. The result of these actions is a much greater apparent reduction in speed of the single-grain-layer plates over the multigrain-layer plates for the long wave-length radiation than for the short-wave-length radiation.

The sensitivity values obtained here (Fig. 6) on the multilayer emulsion at wave-length λ3600A are compared in Table I with values obtained with ordinary emulsions by other workers. The values in Table I give sensitivity in terms of exposure required to produce a density of unity. The values by Jones and Schoen and by Leimbach are expressed as quanta per square micron. If we assume an average grain size of $0.5\mu^2$, then these values would become 200 quanta per grain and 675 quanta per grain, respectively. Considering that the values in Table I are independent determinations for different emulsions of varying speed and obtained by different workers, there is a surprising degree of consistency among the values found.

In Fig. 7, the characteristic curves made at different wave-lengths on the single-grain-layer plates have been replotted as a function of quanta absorbed per average grain. The amount of absorbed light was obtained by multiplication of the incident light per grain by the percentage absorption for a single average grain. The absorption by a single grain was obtained from experimental work carried out by Slade and Toy[15] on absorption of light by thin sheets of silver bromide. In their work, Slade and Toy determined the absorption of thin layers of pure silver bromide with high accuracy (±3 percent). From their data, the curve in Fig. 8 was derived for the percentage absorption of a silver bromide layer of thickness 0.55×10^{-4} cm, which corresponds to the thickness of the average grain of the emulsion under test. Values cannot be used directly from this curve because the grains of the emulsion are spherical, and the absorption by a spherical body of diameter d is less than that

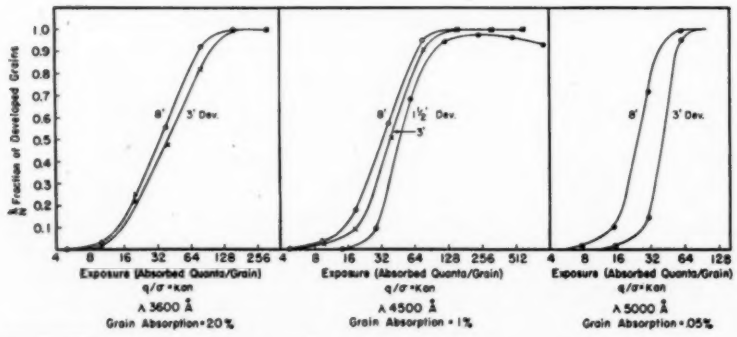


FIG. 7. Characteristic curves on single-grain-layer plates for different wave-lengths.

[15] R. E. Slade, and F. C. Toy, Proc. Roy. Soc. 97, 181 (1920).

TABLE I.

Observer	Incident exposure	Wave-length (A)
Helmick[a]	555 quanta/grain	3650
Jones and Schoen[b]	400 quanta/μ^2	4200
Leimbach[c]	1350 quanta/μ^2	4500
Eggert and Noddack[d]	280 quanta/grain	4028
Webb (present work)	550 quanta/grain	3600

[a] P. Helmick, J. Opt. Soc. Am. and Rev. Sci. Inst. **9**, 521 (1924).
[b] L. A. Jones and A. L. Schoen, J. Opt. Soc. Am. and Rev. Sci. Instr. **7**, 213 (1923).
[c] G. Leimbach, Zeits. wiss. Phot. **7**, 157 (1909).
[d] J. Eggert and W. Noddack, Sitz. Preuss. Akad. Wiss. **39**, 631 (1921).

for a flat disk of thickness d in accordance with the relationship,

$$\frac{\text{absorption of sphere}}{\text{absorption of disk}} = \frac{1-\frac{2}{\mu^2 d^2}\left[1-e^{-\mu d}(1-\mu d)\right]}{1-e^{-\mu d}}, \quad (6)$$

in which μ is the extinction coefficient. This relationship approaches the limiting value $\frac{2}{3}$ as μd approaches zero. The value of μd for the wave-length region considered here varies from 0.3685 at λ3600A to near zero at λ5000A, and over this range of values of μd, the relationship (6) varies from 0.703 to 0.666. Accordingly, the approximation was made of taking the absorption by the spherical grains of diameter d to be two thirds of that for a plain layer of thickness d for all wave-lengths. The curves of Fig. 7 for wave-lengths λ3600A and λ4500A were plotted using the absorption values for pure silver bromide. Although the present emulsion contains a small amount of iodide, of the order of 1 percent, the absorption of the grains at wave-lengths λ3600A and λ4500A would not be greatly influenced by the iodide content. Using the data of Slade and Toy and taking into account the shape of the grains, the absorption of the individual grains of average diameter 0.55×10^{-4} cm is found to be 20 percent at λ3600A and 1 percent at λ4500A. The relative absorption values at these two wave-lengths are seen to be in the ratio of 20:1. From the relative sensitivities of the emulsion at these two wave-lengths, as shown by the separation of the 3-minute development curves of Fig. 6, the factor amounts to 22 in close agreement with the relative absorption values. For the curve made at wave-length λ5000A (Fig. 7), there are no reliable absorption measurements on silver bromide available for independently placing this curve on an absolute absorbed exposure basis. What has been done here is to assume that the relative sensitivity at this wave-length is a measure of the relative absorption. Since the curve for λ5000A is shifted by a factor of 22.4 from the λ4500A curve (Fig. 6), it was assumed that the absorption at this wave-length is approximately 20-fold less than at λ4500A. Thus, the absorption of the average grain at wave-length λ5000A was taken to be one twentieth that at λ4500A, or 0.05 percent. Thus, the three sets of curves plotted in Fig. 7 were placed on an absolute absorbed exposure basis, using the absorption values per average grain shown in Table II. In using the relative sensitivity as the basis for absorption at λ5000A, we realize that we are begging the question, but it is the only way open at present unless independent absorption measurements are made. With the extremely low absorption values of silver bromide at λ5000A, it is doubtful whether a more reliable figure applicable to the case of emulsion grains can be obtained without very elaborate means.

TABLE II. Absorption for average grain.

λ, A	Absorption K %	Basis
3600	20	Slade and Toy's work on pure silver bromide
4500	1	Slade and Toy's work on pure silver bromide
5000	0.05	Relative sensitivity data

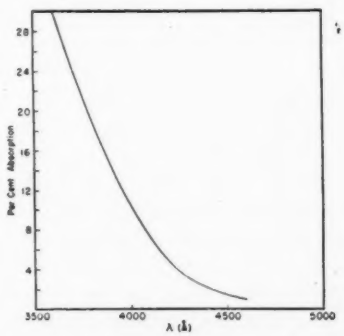


FIG. 8. Spectral-absorption curve for Ag-Br layer of thickness 0.55×10^{-4} cm.

When the characteristic curves are plotted on an absolute basis of absorbed exposure, as shown in Fig. 7, the sensitivity of the grains, as indicated by the exposure values in the toe of the curves, exhibits a high degree of consistency for all three wave-lengths. All of the curves start to rise in the neighborhood of 8 quanta per grain.

Number of Quanta Required to Form the Latent Image

The curves of Fig. 7 show the number of quanta absorbed by an average grain for exposures at different wave-lengths. It cannot be supposed, however, that all quanta absorbed by a grain are actually utilized in the formation of the latent-image speck which initiates development. Some of the quanta may go to forming internal specks of silver or in building up small specks of silver haphazardly distributed in the grain, neither of which necessarily contributes to that single speck of silver which first becomes large enough to initiate development of that grain. Hence, the number of quanta absorbed by a grain in reaching the developable state, even though known exactly for an individual grain, does not necessarily represent the number of quanta producing the latent image in that grain. The number of quanta absorbed would, of course, set an upper limit to the number of quanta forming the latent image but would not uniquely determine this quantity.

There is another well-known method[4,11,14] by which information can be obtained on the number of quanta actually utilized in latent-image formation on a single photographic grain and that is by the use of statistical analysis. We now wish to apply this method to the experimental results obtained in the present study.

It was seen in the case of the experiment with alpha-particle exposure described earlier that the experimental value of fractions of exposed grains rendered developable by known incident exposures of alpha-particles was accurately in agreement with the theoretical formula, (3), with r equal to unity, derived from the Poisson law. Not only did the shape of the experimental exposure curve agree with the theoretical curve (Fig. 3) but also the absolute exposure values in terms of alpha-particles per grain as determined experimentally were found to agree closely with the theoretical values. This agreement in a case for which it is reasonably certain that a grain is made developable by the action of a single incident particle ($r=1$) gives considerable support for the soundness of this method of analysis. Furthermore, the results of this calibrating experiment have ruled out of account a number of spurious effects which might be imagined to enter into the present experiments to vitiate the results, such as variation in grain size, errors in counting grains, non-uniformity of grain coatings, statistical errors, etc.

By the use of the experiment with alpha-particle exposure as a guide, a similar analysis of the light-exposure experiments is also possible by means of the theoretical curves of Fig. 4. These curves were drawn from an equation of the type (3), namely,

$$\frac{k}{N}=\left[1-e^{-q}\sum_{j=0}^{r-1}\frac{q^j}{j!}\right], \tag{7}$$

in which q represents the average number of quanta photo-efficiently absorbed by a grain, and r represents the least number of such quanta required to form the developable latent image. All grains photo-efficiently absorbing r, or more, quanta will be developable grains. For convenience of analysis, the quantity q can be broken up into a number of separate factors,** as

$$q=\sigma\kappa a n, \tag{8}$$

in which n represents the number of quanta per square centimeter incident on the single-grain-layer plate, a is the area of an average grain (here $0.275\mu^2$), κ is the absorption coefficient for a single average grain (Table II), and σ represents the fraction of quanta absorbed that are actually utilized by the grain in forming the latent image. The factor σ may correspond to a fractional area of the grain (all in one piece and surrounding the sensitivity center) as an effective receptor area for the absorbed quanta, or it may correspond simply to the photographically effective fraction of all the quanta that are absorbed at random over the grain. In the present analysis we cannot distinguish between these two interpretations.

** In Eq. (3) for alpha-particle exposure, the factors σ and κ were assumed to have the value unity.

The curves of Fig. 4 represent Eq. (7) with r values of 1, 2, 3, 4, 10, and 100. The experimental curves of Fig. 7 are to be compared with the theoretical curves of Fig. 4 for shape in order to determine which r value gives the best representation of the experimental points. As may be seen by inspection, the slopes of the curves of Fig. 7 differ considerably from one another for different wave-lengths. Accordingly, different r values would be obtained for each of these sets of curves. However, it seems more logical to believe that the same number of quanta are required to form the latent image independently of wave-length, a view that is strongly supported by the equality of shapes of the reciprocity-law failure curves at different wave-lengths.[16] This raises the question then as to which one of the curves of Fig. 7 has a shape corresponding to the significant value of the quantity r. To throw light on this question, we need to consider the exposure problem a little more carefully from the statistical standpoint.

If Eq. (7) is to give a true answer to the question of the number of discrete events, r, required in production of the latent image, a number of conditions must be satisfied. There should be no variation in size among the grains of the emulsion under test, there should be no variation in the absorption factor among grains, there should be no variation in the utilization factor σ for the quanta absorbed by a single grain, and each grain should have an absolute threshold sensitivity, r, the same for every grain. In other words, every grain should be absolutely identical with every other grain relative to the mechanism of reception of quanta and subsequent utilization of these quanta in acquiring developability. Under these conditions, Formula (7) would apply rigorously. For a definite value of r, an exposure q, calculated from the relation,

$$q = \sigma\kappa a n,$$

would determine a unique value of k/N. Under these ideal conditions, the curve of k/N *versus* q would have a unique slope characteristic of the r value. A variation of any of the factors (grain size a, absorption κ, σ, r) among the grains would tend to spread the curve so as to decrease the slope of the curve.

[16] J. H. Webb, J. Opt. Soc. Am. **23**, 316 (1933).

On the basis of this analysis it was concluded that the most reliable curve of Fig. 7 for a determination of the r value is the steepest curve. If the curve, for wave-length λ5000A and development time 3 minutes, is chosen for this purpose and compared with the theoretical curves of Fig. 4, it is found that the theoretical curve drawn for $r=10$ gives the closest approximation to the experimental curve. If, now, the q value of the theoretical curve ($r=10$) of Fig. 4 be compared with the q/σ value of the experimental (λ5000A, 3 minutes' development) curve of Fig. 7 for equal k/N values, the value of σ can be calculated. For example, at the k/N value of 0.5, q is found to be 9.6 and q/σ is found to be 40. Therefore,

$$q \div q/\sigma = \sigma = 9.6/40 = 0.24.$$

This means that of the q quanta absorbed by a grain a fraction 0.24, in this case, is actually utilized in formation of the latent image.

It might be argued that the characteristic curve for λ5000A chosen for the determination of the r value is still not the steepest curve attainable and that therefore the r value obtained may be too low. This is a legitimate criticism and therefore the r value is merely a lower limit value. However, the quanta absorbed per grain, as shown for the curves of Fig. 7, were determined by experiment and therefore set an upper limit to the value of r. In the cases of the λ3600A and λ4500A curves, the absorbed exposure values are accurately known and accordingly fix fairly closely the upper limit for the actual number of quanta utilized by a grain in formation of the latent image. Thus, it may be concluded that the r value determined from the shape of the curve sets a lower limit for the number of quanta utilized in forming the latent image, whereas the measured values of absorbed quanta shown on the abscissa axis give an upper limit. In the case of the λ5000A curve (3 minutes' development) of Fig. 7, it appears that the true value of r must lie between the limits of 10 and 16*** in the toe of the curve and between 10 and

*** The value 16 instead of 8 was chosen here because of statistical fluctuations in numbers as small as 8. Thus, if the average number of absorbed quanta per grain as determined experimentally is 8, the relative probability of 12 quanta being absorbed is about one third. Hence, the higher value 16 was selected as a safer upper limit.

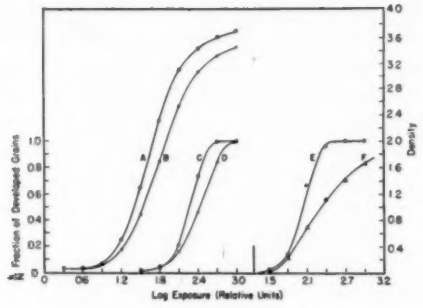


FIG. 9. Reciprocity-law failure for white light; developed 3 min., D–16 at 68°F. *A*, *B* = multilayer emulsion, *C*, *D*, *E*, *F* = single-layer emulsion,

Curve	Rel. intensity	$T_{max.}$
A, *C*, *E*	1	8 sec.
B, *D*	1/1024	2 hr. 16 min. 32 sec.
F	1/16,384	36 hr. 24 min. 32 sec.

64 in the shoulder of the curve. The r values for the curves (λ3600A and λ4500A) of Fig. 7 cannot be uniquely determined from the shapes of these curves. On the basis of this argument, the resulting curves in these cases must be a composite of several curves, all of the same r value, but spread apart because of variation of the σ factor among the grains. The point of view that has been adopted here is that the steepest attainable experimental curve for a single-grain-layer emulsion is the nearest approach to the ideal shape for determination of the r value in latent-image formation. From the statistical standpoint, the ideal condition can be approached but cannot be surpassed. As to why the experimental curves actually become steeper at the longer wavelengths in these experiments, it is not at present clear. Tentatively, we believe that the σ-factor exhibits less variation in magnitude among the grains for very weakly absorbed light. It is proposed to study this effect further by making experiments with other emulsions.

Variation of r with Development Time

As pointed out earlier, the characteristic curves for different development times vary somewhat in shape and are shifted toward higher exposure values with shorter development time, indicating lower sensitivity. This, of course, is to be expected, since the less strongly exposed grains would be brought up to developability only by increasing the development time. The change in shape of the curve with development time seems to consist in the main of a steepening of the curve with shorter development time. Two of the sets of curves in Figs. 5, 6, and 7 show this effect, and other data not included here have shown this effect to be quite generally true for the present emulsion. If it is assumed that a shorter development time will develop only those grains which possess large, well-grown, latent-image specks, this effect is understandable. Only those grains for which the latent-image speck is large, and therefore built by a larger number of quanta (high r value), develop at all with short development. Since for larger r, the curves should become steeper, it is to be expected that short development time should lead to steeper curves. Herein is an example where experiments with multilayer emulsions would almost certainly lead to error, because it is always true, insofar as the author is aware, that multilayer emulsions give steeper curves with increased development time.

Reciprocity-Law Failure Measurements

In Fig. 9 are shown characteristic curves on the single-grain-layer plates made with exposures at widely varying intensity levels. These exposures were made to unfiltered tungsten light and are time-scale curves. The two curves, *C* and *D*, were made at intensities of light in the ratio of 1:1/1024, and the times of the maximum exposure steps were 8 seconds and 8096 seconds, respectively. The two curves, *E* and *F*, on the right were made at intensities in the ratio of 1:1/16,384, and the times of the maximum exposure steps in these cases were 8 seconds and 131,072 seconds, respectively.

As may be seen from the curves of Fig. 9, there is a considerable drop in sensitivity of the single-grain-layer emulsions as the intensity level of the exposure is lowered. Also, there is a marked drop in the slopes of these curves at lowered intensity. In fact, the change of shape of these curves with lowered intensity is perhaps the most important feature of this study. It would seem to indicate that the grains responsible for the upper portions of these curves are more subject to reciprocity-law failure than are the grains of the lower part of the curves. Since this effect

is probably connected with the stability of the latent-image specks during the very early stages of the exposure, it indicates that the original sensitivity specks in different grains may differ widely. Before these experimental results were obtained, it was anticipated that the characteristic curves at lowered intensity would approach in shape that of the one-quantum-hit formula ($r=1$) shown in Fig. 4. This anticipation was based on the postulate that a grain acquires a stable sublatent image only when a burst of say 2, 3, or 4 quanta is received within a certain prescribed short time interval. At very low intensities, as was shown theoretically by Berg[17] and also by Silberstein,[18] these bursts will occur in time with a probable distribution just as individual events do in a single-hit Poisson distribution. However, the trend of the curves in Fig. 9 with lowered intensity does not seem to be moving in the direction of the single-hit type of curve. Accordingly, to explain the results, we must assume in this case variation among the grains in their response to low-intensity exposure.

In addition to the curves obtained on the single-grain-layer emulsions, two characteristic curves, *A* and *B*, are shown for the multilayer emulsion. These curves are plotted as density *versus* log exposure, and the emulsion is composed of the same grains as were used for the single-grain-layer plates. Curve *A* had the same exposure conditions as curve *C*, and curve *B* had the same exposure conditions as curve *D*. Curves *A* and *B* are correctly placed on the abscissa axis according to exposure values, and thus show the relative sensitivities of the multigrain-layer emulsions relative to the single-grain-layer emulsion. The two curves, *A* and *B*, show the extent of the reciprocity-law failure on the regular multilayer-grain emulsion relative to that for the single-grain-layer emulsion over the intensity range 1:1/1024.

ACKNOWLEDGMENT

In conclusion, the author wishes to express his appreciation to Mr. R. P. Loveland, of these Laboratories, under whose direction the single-grain-layer plates used in these experiments were prepared. The excellence of these plates contributed considerably to the execution of this work.

[17] W. F. Berg, Phil. Mag. (Series 7) **36**, 337 (1945).
[18] L. Silberstein, Phil. Mag. (Series 7) **36**, 319 (1945).

Physics Abstracts

For many years the dues of the American Physical Society have included a subscription to Physics Abstracts, published by the Institute of Electrical Engineers of London. There has been, however, no convenient method by which persons not members of the A.P.S. could receive this publication. The Optical Society of America has now instituted an arrangement under which any O.S.A. member can receive Physics Abstracts at an annual subscription price of $4.50. All correspondence in this connection should be addressed to the American Institute of Physics, 57 East 55th Street, New York 22, New York.

Performance of an Optical System as the Sum of the Contributions of Its Construction Elements*

M. HERZBERGER
Kodak Research Laboratories, Eastman Kodak Company, Rochester, New York

(Received October 31, 1947)

The investigations of Cruikshank have brought to the attention of American scientists different methods of analyzing the optical image, methods which were also suggested by A. Kerber and the author. However, all these methods are, for the most part, restricted to the analysis of meridional rays.

In previous papers the author has shown the importance of the analysis of the diapoint configuration for the optical image. He here gives a simple formula to show how the diapoint deviations can be considered as the sum of errors originating at the single surfaces. In this way, and with the help of a series of diagrams, the lens designer can obtain an accurate picture of the effect of the single surface on all the rays traversing the optical system. Diagrams for two photographic lenses are included.

BECAUSE of the work of F. Cruikshank and H. L. M'Auley,[1,2] optical designers in this country have become interested in methods of differential analysis of the image errors. This is a subject on which there exists a sizable but neglected European literature which goes as far back as L. Seidel.[3]

Notable contributions in this field have been made by A. Kerber,[4] K. Schwarzschild,[5] F. Nusl,[6] and T. Smith;[7,8] the author[9] published in 1927 relatively simple formulas which permit these methods to be applied to skew rays (compare K. Schwarzschild[5]).

In all these cases the problem is to analyze the deviation of the intersection points of a bundle of rays from a reference point in the image plane, i.e., in a plane perpendicular to the optical axis.

In recent years the author has frequently commented[10,11] upon the importance of studying the configuration of diapoints, i.e., the configuration of the intersection points of the skew rays in the image space with the meridian plane (the plane through the object point and the axis of the system). In a previous paper[11] a new image error theory based upon this diapoint configuration was outlined. It seemed worth while, therefore, to analyze the deviations of the diapoints, and it was not surprising to find that such an analysis gives a deeper insight into the performance of optical systems than previous methods have given.

The Staeble-Lihotzky[12,13] condition gives a measure of the coma for points near the axis of an optical system. We found this condition to be a special case of a general formula for the asymmetry of an off-axis image point.

In general, the diapoints of a given object point form a two-dimensional manifold of points in the meridian plane. If this manifold degenerates to a curved line it is said to be free from *deformation* errors; if the line is straight, the image is free from *asymmetry* errors, and if all the diapoints degenerate to a point, the image is *sharp*.

* Communication No. 1165 from the Kodak Research Laboratories.

[1] F. Cruikshank, "A system of transfer coefficients for use in the design of lens systems," Proc. Phys. Soc. **57**, 350, 362, 414, 426, 430 (1945).

[2] H. L. M'Auley, "A transfer method for deriving the effect on the image," Proc. Phys. Soc. **57**, 435 (1945).

[3] L. Seidel, "Trigonometrie der Formel für den allgemeinen Fall," Münchener Sitzber. (2) 263 (1866).

[4] A. Kerber, *Beiträge zur Dioptrik* (Leipzig, 1895/99).

[5] K. Schwarzschild, *Über differenzen Formeln zur Durchrechnung optischer Systeme*, (Göttingen, 1907).

[6] F. Nusl, "Über allgemeine differenzen Formeln," Bull. Int. Acad. Sci. de Bohême, **12**, 84 (1907).

[7] T. Smith, "The general form of the Smith-Helmholtz equation," Trans. Opt. Soc. **31**, 241 (1929).

[8] T. Smith, "Variational formulae in optics," Proc. Phys. Soc. **57**, 558 (1945).

[9] M. Herzberger, "Über die Durchrechnung von Strahlen durch optische Systeme," Zeits. f. Physik **43**, 750 (1927).

[10] M. Herzberger, "A new theory of optical image formation," J. Opt. Soc. Am. **26**, 197 (1936).

[11] M. Herzberger, "The limitations of optical image formation," Ann. N. Y. Acad. Sci. **48**, 1 (1946).

[12] F. Staeble, "Über den Zusammenhang von Koma und Sinusbedingung," Zeits. f. Instrumenten. **27**, 241 (1907).

[13] E. Lihotzky, "Verallgemeinerung der Abbe'schen Sinusbedingung," Wiener Sitzber. **128**, 85 (1919).

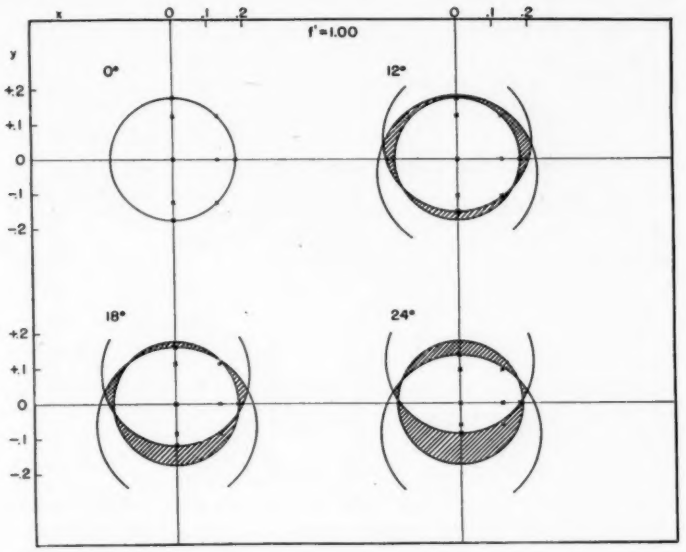


FIG. 1. Vignetting diagram for an optical system.

In general, the diapoint of an object point is given by two data for which we can take its *longitudinal* aberration, i.e., its distance z from a suitable origin and its *diamagnification*, m. For an infinite object point the latter is replaced by the diafocal length (compare M. Herzberger[10]).

It is obvious that the diamagnification is the product of the diamagnifications of the single surfaces. Thus, if the logarithms of the diamagnifications for the rays from an object point traced through the system are plotted, the plot will be the sum of plots showing the diamagnification for each single surface.

The analysis of the longitudinal aberration is done in the following way: Two Cartesian coordinate systems are assumed, one in the object space, and one in the image space, with the origins at the centers of the first and the last surface, respectively. The z and z' axes coincide with the optical axis, and the $y(y')$ axis may lie in the meridian plane. Then, for the coordinates y', z' of the diapoint, it is found that

$$\begin{aligned} y' &= my, \\ z' &= mz + S, \end{aligned} \tag{1}$$

if y, z are the coordinates of a finite object point. Correspondingly, if the object point is at infinity it is found that

$$\begin{aligned} y' &= f\frac{\eta}{\zeta} \\ z' &= f + S, \end{aligned} \tag{2}$$

where ξ, η, and ζ are the direction cosines of the entering ray (multiplied by the refractive index of object space) and f is the diafocal length. Analysis of the explicit formula will show that S (the difference between the changes of the back focus and the focal length) is a linear function of the center distances c_i, thus,

$$S = \sum \alpha_i c_i. \tag{3}$$

If S equals zero, y' and z' are proportional, i.e., the diapoints lie on a straight line, and the image rays are symmetrical around this straight line. S therefore gives a measure of the asymmetry of the optical system. This is the generalization of Staeble-Lihotzky's condition for the off-axis points.

The suggestion offered here is to plot, for a suitable number of chosen rays from an object point, the logarithm of the diamagnification, m, and the asymmetry error, S. The two plots can be considered as a sum of plots, one corresponding to the contributions of the single surfaces, and

the other corresponding to the contributions of the center distances.

As independent coordinates we shall choose data which define the point of incidence in the entrance pupil.

Derivation of the Calculation Formulas

For a single spherical surface the connection between point and diapoint is given by the relations

$$\begin{aligned} y_1' &= m_1 y_1, \\ z' &= m_1 z_1, \end{aligned} \tag{4}$$

when the y–z plane is the meridian plane, that is, the plane which contains both the axis and the object point.

For the computation of m, see a previous paper by Herzberger.[14] Here, the formulas are given without derivation. The object point is given by y and z, and the direction of the object ray by the optical direction cosines ξ, η, and ζ (by "optical" cosines we mean the cosines multiplied by the refractive index, μ).

Abbreviating, we have

$$\begin{aligned} \tau &= y\eta + z\zeta, \\ Q^2 &= \mu^2(r^2 - x^2 - y^2) + \tau^2, \\ Q'^2 &= \mu'^2(r^2 - x^2 - y^2) + \tau^2, \end{aligned} \tag{5}$$

where r is the radius of the sphere, and μ, μ' are the refractive indices of the first and second medium. We compute

$$\begin{aligned} \lambda &= \frac{1}{\mu^2}(Q - \tau), \\ \Phi &= \frac{1}{r^2}(Q' - Q), \end{aligned} \tag{6}$$

and find the diamagnification from

$$m = (1)/(1 + \lambda\Phi). \tag{7}$$

For the refraction at the second surface there are equivalent formulas, except that the origin must be shifted from one center to the next. That leads to

$$\begin{aligned} y_2' &= m_2 y_2, \quad \text{and} \quad y_2 = y_1' \\ z_2' &= m_2 z_2, \qquad\quad z_2 = z_1' - c_1, \end{aligned} \tag{8}$$

where c_1 is the distance from the first to the

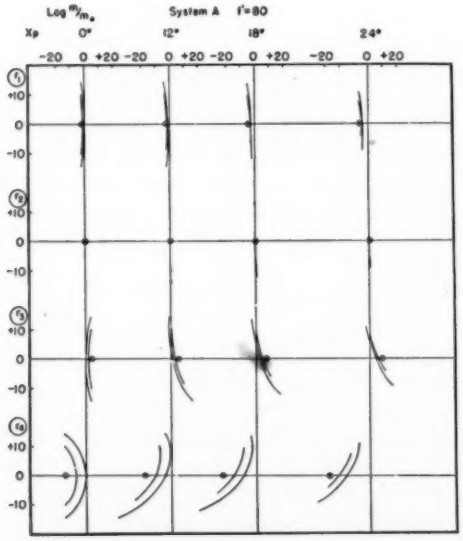


FIG. 2. Plot of diamagnification contribution of four surfaces.

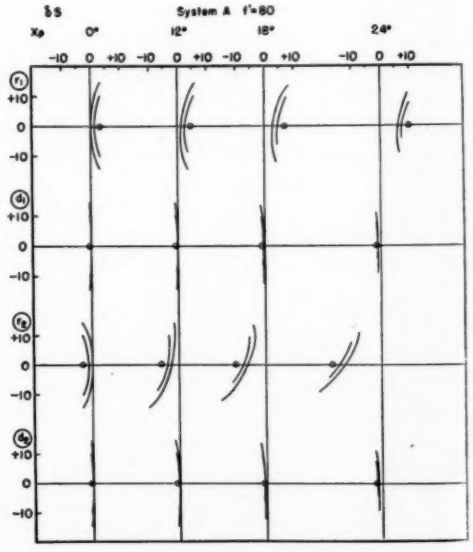


FIG. 3. Plot of asymmetry contribution for two radii and two thicknesses.

[14] M. Herzberger, "A direct image error theory," Quart. App. Math. 1, 69 (1943).

second center. Abbreviating, we find that

$$m_k m_{k-1} \cdots m_\lambda = M_\lambda. \tag{9}$$

One obtains finally

$$\begin{aligned} y' &= M_1 y_1, \\ z' &= M_1 z_1 + S, \end{aligned} \tag{10}$$

where

$$S = M_2 c_1 + M_3 c_2 + \cdots c_{k-1} = \sum_1^{k-1} M_{v+1} c_v. \tag{11}$$

These are the formulas which we wanted to derive. This quantity, S, may be interpreted as the sum of the center distances projected to the image space.

Selection of Rays and Plotting Procedure

The selection of the rays to be traced should give as clear a picture as possible of the behavior of all the rays which pass through the system. The following procedure is suggested.

Figure 1 shows, for four angles (a beam parallel to the axis, at two zonal angles and a beam from the corner of the field), the vignetted entrance pupil. This is a circle on the axis, and it is reduced for off-axis points because the first and last surfaces cut off some of the rays. The y axis in Fig. 1 ($x=0$) characterizes the meridional rays, the only ones considered by most designers. The meridian plane is a plane of symmetry for the image formation. It is therefore sufficient to study the rays going through the right half of the pupil.

In general, we trace five meridional rays for each angle distributed over the aperture of the system and plot m and S against the aperture height. (It may be necessary to remark here that for meridional rays the diapoint becomes the sagittal image point.) This is one of the series of curves plotted in Figs. 2 and 3.

For the second curve we trace three or more skew rays which intersect the pupil in a line $x=$ constant of medium skewness, and finally we trace a single ray at maximum skewness through the system. From these calculations it is possible to obtain values of m and S for all the rays passing through the entrance pupil by using simple interpolation methods. This ties up the curves

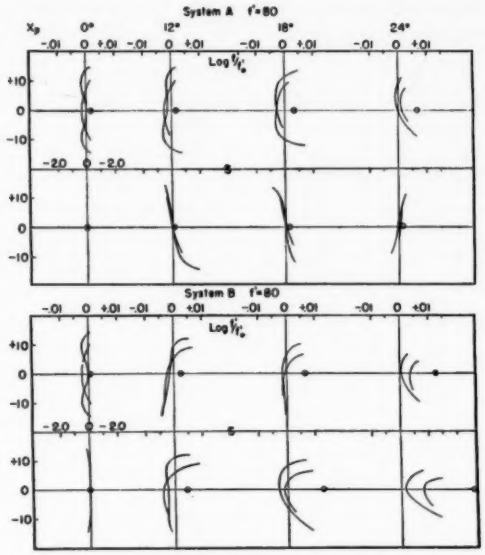


FIG. 4. Diamagnification and asymmetry error for one system: System A (upper curves) is the final form, and system B (lower curve) is a form before final corection.

thus obtained with the spot diagrams shown by Herzberger.[15]

Additional Remarks

In the preceding paragraphs we split up the asymmetry errors, S, into the contribution of the center distances, c_i. This is the simplest way to interpret the asymmetry errors, and the finding of this balance is the chief task of the optical designer. However, the center distances are not physical construction elements of an optical system, and, as such, are not easy to visualize. The designer would prefer to use curvatures $(1/r_i)$ and thicknesses (t_i) for this purpose. Since

$$c_i = t_i + r_{i+1} - r_i, \tag{12}$$

where r_i and r_{i+1} are the preceding and following radii, respectively, we find from (11) that

$$S = \sum M_{v+1} t_v + \sum (m_v - 1) r_v M_{v+1}. \tag{13}$$

Formula (13) makes it appear that S is proportional to the radii and would become infinite if r tends to infinity. This is, however, not the case.

[15] M. Herzberger, "Light distribution in the optical image," J. Opt. Soc. Am. **37**, 485 (1947).

Formulas (5) to (7) permit us to derive the following formula:

$$(m_v-1)r_v=-\frac{(Q'-Q)}{1+\gamma\Phi}\gamma, \tag{14}$$

which, for infinite radius, converges to

$$\lim_{r\to\infty}(m-1)r\to-\lambda(\mu'-\mu), \tag{15}$$

which is finite, since λ, the distance from object point to surface, is finite. By use of the procedure already described for (15), the method developed can be applied to systems which have plane surfaces or surfaces with long radii.

Figure 4 shows how these methods can be used by the optical designer to correct the system by reducing excessive aberrations due to a single surface.

Another interesting result is that the balance between different errors seems to be connected with the type of system so that, for the fine correction of an optical system, the designer is limited to correcting those errors which are not inherent in the type of system.

Book Review

Reading and Visual Fatigue

LEONARD CARMICHAEL AND WALTER F. DEARBORN. Pp. 483+xiv. Houghton Mifflin Company, Boston, 1947. Price $5.00.

This book is a report of the authors' study of reading performance in which they found that under the conditions used there was no "work decrement," no permanent change in pattern of eye movements, no signs of irritability or nervousness as result of reading of fiction or non-fiction for six consecutive hours. Although the authors were loathe to select an explicit definition of fatigue from the literature or formulate one of their own, they concluded that their findings indicated no fatigue at all. A large fraction of the twelve chapters in the book is given over to presentations of the literature on the recording of eye movement, blink rate, the electro-oculogram, etc. This they did in a most excellent fashion. The chief merit of the book then lies in the authors' having done this very thing, not in having presented something new about fatigue. It is well known that (using the very same assumptions as to what fatigue is) some individuals do become fatigued after a very short period of reading, whereas others are inveterate readers and, day after day, spend many hours at it. The authors did show that reading projected microfilm is not likely to lead to more difficulty ("fatigue") than reading book print under the same conditions. Apparently the factor of motivation was sufficiently controlled in the subjects either by extrinsic rewards or by interest in the experiment itself, that it was not a material variable, leading to early fatigue in some cases.

S. HOWARD BARTLEY
Michigan State College

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 38, NUMBER 4 APRIL, 1948

Visibility on Cathode-Ray Tube Screens: Signals on a P7 Screen Seen at Different Intervals after Excitation*

A. L. SWEET AND N. R. BARTLETT
Department of Psychology, The Johns Hopkins University, Baltimore, Maryland
(Received October 22, 1947)

When signal images appear on the phosphorescent screen of a cathode-ray tube, they persist for some time before disappearing from view. This study investigated the visibility of signal images on one type of screen at different intervals from excitation to final disappearance.

Data were furnished by five observers viewing signal images on a P-7 screen of a PPI cathode-ray tube. It was found that the tube bias, which sets the brightness intensity of the rotating sweep, affects both the level of visibility and the duration of visibility. Increasing the bias, and hence dimming the scope, makes the signal traces visible over a longer period of time. Greatest visibility, at intervals longer than seven seconds after excitation, was obtained at the highest bias used. For more recent signals, optimum visibility results from using a lower bias but which yet presents a dim sweep. At high biases, a signal is at maximum visibility at the instant when it is excited by the sweep. At low biases, a signal is most visible immediately after its excitation. These patterns of visibility can be explained in terms of the contrast of the brightness level of the signal image with that of the scope background and by the lessened differential sensitivity of the eye at low intensities.

INTRODUCTION

IN radar operation, information about the location of objects may be furnished by presenting a light on a screen. With luminescent screens, this light is phosphorescent radiation resulting from the bombardment of a part of the screen by electrons. The screen brightens up at the site of excitation forming a small light area (pip) which stands out from the rest of the screen. The light from this pip may then persist for some time afterwards before it finally disappears, and so, for some time after excitation, the pip remains visible. The degree of persistence varies from one type of luminescent screen to another. The P7 screen investigated in this study has, in comparison with other type screens for cathode-ray tubes, a medium to long persistence so that the radar operator may see a pip for several seconds after its original excitation.

The time factor in phosphorescence, then, is one of the conditions which govern the visual detection of signals. Most studies on phosphorescence, however, have been concerned with such purely physical problems as the decay curve of the radiant energy and spectral composition of the emitted light.[1] Comparatively little investigation has been made of the effect of phosphorescent decay of light on visibility.[2]

There are several pertinent questions concerning visibility on radar scopes which relate to phosphorescence. For example, at what time should the operator look at the signal for greatest visibility? He can, for instance, "ride" the sweep with his eyes and attempt to perceive a signal image the instant it is being excited; or he may follow the rotating sweep with his eyes fixed at some distance behind in order to detect pips which are already in a state of decay. Should he "ride" the sweep? If not, by how much should he follow it?

Some further related questions are: what duration of visibility can be expected for signals of different strengths? How is the visual persistence of signals affected by the brightness level of the scope? What scope brightness gives optimum visibility for signals of different ages?

This experiment was designed to obtain data on these problems. Of course, the data depend on the specific type of luminescent screen used. In this investigation we were particularly concerned

* This research was carried out under Contract N5ori-166, Task Order I, between Special Devices Center, Office of Naval Research, and The Johns Hopkins University. This is Report 166-I-35 under that contract. The authors are indebted to Mr. Ellwood King, Electrical Engineering Laboratory, for the photometric measurements used in the report and for other engineering assistance in carrying out the study.

[1] H. W. Leverenz, R.C.A. Rev. **7**, 199 (1946).
[2] S. B. Williams and N. R. Bartlett, "Visibility on cathode-ray screens: Problems and methods" (in preparation).

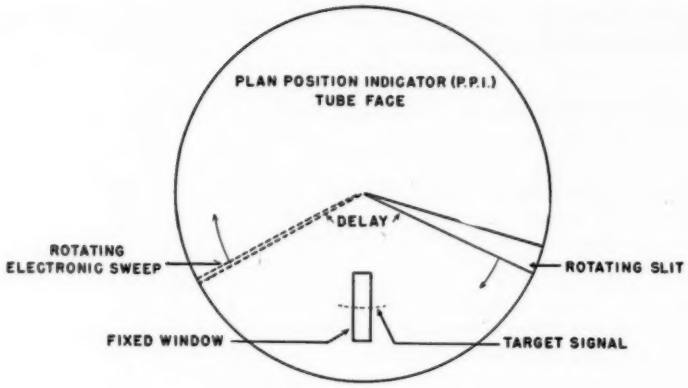


FIG. 1. General scheme of apparatus used for studying visibility of signals at different exposures with varying amounts of delay following electronic excitation.

with radar applications of luminescent screens. On the basis of its wide usage with radar, therefore, the P7 screen was studied.

APPARATUS AND PROCEDURE

The P7 Screen

Signals whose visibility was to be measured were exposed on a P7 luminescent screen of a PPI scope with an effective viewing diameter of approximately 6 inches. This screen produces a blue fluorescent trace with a yellow phosphorescence. It is of medium to long persistence: in 3 seconds, it drops to approximately 1 percent of its initial brightness.[1] It is widely used for radar cathode-ray tubes.

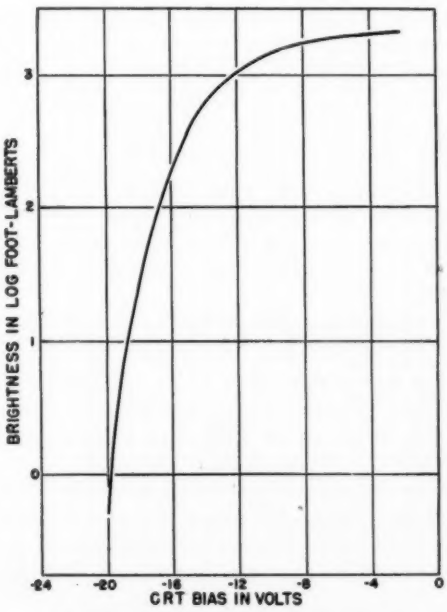


FIG. 2. Peak brightness of the PPI sweep at different CRT biases.

Characteristics of the Signal

The characteristics of the signal presented on the PPI scope were controlled by means of a target generator. The strength of the signal could be varied by an attenuator unit in ½-decibel steps over a range of 0–105 db below a reference voltage. The intensity of the signal is thus measured by a logarithmic scale of attenuation: the smaller the number of decibels, the greater the signal strength.

The reference voltage was defined by a visual calibration procedure performed at the beginning of every experimental session in order to insure the compatibility of the data collected on different days. With the CRT bias set at a voltage where the sweep was just detectable, the attenuator scale was set to read 28-db signal attenuation. The video gain, which controlled the voltage of the signal, was then varied till an observer, who was repeatedly used for all calibrations, reported a just visible signal. This procedure fixed a constant reference voltage which at 0 attenuation was of the order of 8 volts.

The pulse repetition rate was always 600 per second and the length of the signal pulse was two microseconds.

One signal was exposed to an observer. This signal was a 20° arc located at due south from the center of the PPI scope three-quarters of the way

out from the center to the periphery. A fixation point of red light, just above the window exposing the signal, served to orient the observer and to insure foveal vision. The PPI scope was viewed through a standard hood, thus limiting the field of vision to the tube face and maintaining a constant distance of 12 inches between the observer's eyes and the scope. A sufficient time was allowed for the observers to view the scope for adaptation to the scope brightness before visual thresholds were measured.

Apparatus for Controlling Times of Delay

Figure 1 illustrates the apparatus by which varying times of delay between excitation of the signal image and its exposure to the observer were obtained. Actually, there are two masks, consisting of black paper, represented in this diagram. The inner one blacks out the whole tube face except for a narrow fixed window, $\frac{3}{16}''$ wide and $\frac{19}{32}''$ long, through which the signal is visible. Concentric with this inner mask and lying above it is another mask of black paper with a 4° slit cut out with the same width as the fixed window. *This outer mask is part of a disk which rotates at the same speed as the antenna sweep.* Thus the signal image is made visible to the outside only when the rotating slit passes by the fixed window. (At the constant speed of 10 r.p.m. used in this study, this gave an exposure time of 0.067 second.)

Since the rotating sector and the sweep revolve at the same speed, the slit can be placed at any fixed angle with respect to the sweep. This angle is a measure of the time interval between the instant of excitation of the pip by the sweep and the exposure of that pip to the observer. When the rotating slit is made to coincide with the sweep so that the delay angle is zero, the observer sees the image in a state of fluorescence: the signal image is exposed only at the moment of excitation. The rotating sector can also be set so that its leading edge is just behind the sweep: this will expose the signal immediately after fluorescence. Or the rotating sector may have its leading edge several degrees behind the sweep. Thus, by varying this angle between the sweep and the rotating slit, different durations of delay up to 6 seconds could be obtained. For periods of delay greater than 6 seconds, which was the maximum that could be obtained by angular variation at the standard rotational speed of 10 r.p.m., the moving sweep was stopped for the desired number of seconds and then resumed in its rotation.

The following were the times of delay at which thresholds were measured: fluorescence, post-fluorescence estimated to be 0.04 second, and delays of 0.5, 3, 7, 15, 30, 60 seconds. In a second part of the experiment the duration of visibility for a signal of zero attenuation was determined.

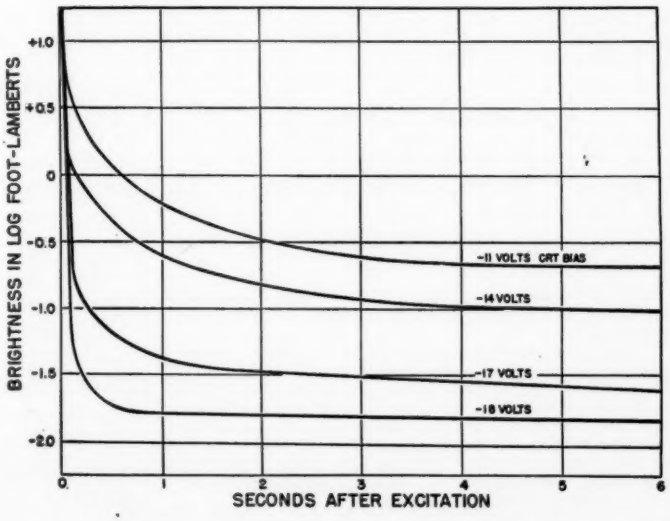


FIG. 3. Decay in brightness of scope background after excitation by sweep at four different CRT biases.

TABLE I. Visibility in decibels of attenuation below reference voltage for different times of delay at four different CRT biases (means of 10 readings: 2 for each of 5 observers).

CRT bias in volts	Duration of delay in seconds												
	0	0.04	0.5	3.0	7.0	15.0	20.9	30.0	33.1	60.0	63.3	141.0	195.8
−11	20.9	24.1	22.7	18.5	9.1	6.0	0						
−14	22.0	26.6	25.0	21.0	15.4	8.0			0				
−17	31.7	31.4	29.1	24.9	21.2	15.8		8.1			0		
−20	30.0	25.2	22.0	21.2	22.0	19.1		19.0		16.0		10	0

CRT Bias and Scope Brightness

A third primary variable controlled in this investigation, the other two being signal strength and time of delay, was cathode-ray tube or CRT bias. This is an electrical control which varies the brightness of the sweep and hence the brightness of the scope. Visibility in this experiment was studied under four different CRT biases: 11, 14, 17, and 20 volts. The 11-volt bias is close to the maximum brightness that was available on the PPI scope. The peak brightness of a small area on the sweep at this bias is 1220 foot-lamberts, according to measurements by a photo-multiplier tube technique.[3] The bias of 20 volts is at the other extreme of brightness and furnishes a dim screen where the sweep line is just detectable: its peak brightness is 0.5 foot-lamberts. The peak brightnesses of the sweep at the intermediate biases were 90 foot-lamberts (17-volt bias) and 630 foot-lamberts (14-volt bias). Figure 2 illustrates the relationship between CRT bias and peak brightness of the sweep. And Fig. 3 shows the decay in brightness with time after excitation (up to six seconds) from the peak intensities of four different CRT biases.**

Determination of Thresholds

In the establishment of each threshold, both ascending and descending orders of signal strength were used. That is, on a given trial, the initial intensity setting of the image would be either definitely above threshold or definitely below threshold, and the observer would direct the experimenter to decrease the intensity or increase the intensity till an image resulted which the observer could just barely see.

Observers

Five male college students, of normal vision, served as observers in this experiment. The visibility function for each over the decay interval of each CRT bias was measured twice.

RESULTS

Reliability of Data

A preliminary analysis of the data, made after each observer had been tested once throughout the decay interval for each bias, showed high agreement among observers in the shape and values of the visual decay functions. The data were analyzed again after each subject had repeated the observations. The curves were almost identical with those obtained in the first half of the experiment, indicating both intra-observer as well as inter-observer agreement.

CRT Bias and Duration of Visibility

Table I presents the numerical values for visibility over the decay interval for each of the four CRT biases. Figure 4 shows the main results of the experiment. Each plotted point is the mean of 10 readings, 2 for each of the five observers. From this graph, several facts are evident. These are:

(1) Visibility for immediate images is better at a bias of 17 volts than at any of the others tested. (This bias is about 5 volts less than a bias at which the sweep can be just barely seen.)

(2) However, as the image decays in time, beyond 7 seconds, greatest visibility occurs at the higher bias of 20 volts.

[3] F. Hamburger and E. King, "A recording photometer and its use in studies of cathode ray screen displays" (in preparation).

** The decay interval was limited to six seconds by the photometric procedure used, which was to photograph a point on the scope as the sweep was revolving. This meant, at the constant speed of 10 r.p.m., a peak intensity every six seconds. Moreover, the photo-multiplier tube technique was not adequate for recording the decay in brightness of the dim scope at 20 volts CRT bias. The curve for the bias of 18 volts, however, can be used as an indication of the decay function for the 20-volt bias.

(3) The duration of visibility of images for the strong signals varies widely, depending on the bias. In general, the higher the CRT bias, the longer the visual persistence of the strong signals. At a signal strength of 0-db attenuation, for example, the persistence of visibility at biases of 11, 14, 17, and 20 volts were as follows: 21, 33, 63, and 196 seconds.

(4) The biases of 11, 14, and 17 volts have parallel functions for visual decay. Within this group, as the bias lessens, not only does visibility decrease, but it falls off in time more rapidly. The bias of 20 volts, however, reveals a somewhat different function: after an initial drop-off, the visibility remains constant till about 15 seconds when it starts to decay in a similar fashion to the other biases.

Other findings are indicated in Figs. 5 and 6.

Signal Strength and Duration of Visibility

Figure 5 reverses the coordinates of the previous graph and shows the logarithm of the time of visual persistence as a function of signal strength. This graph demonstrates the fact that the stronger signals (on the left-hand side) are seen for longer periods and again shows that the duration of visibility increases with the bias. The graph also shows that the signal strength of 21 db is a transition point between the strong signals to the left and the weak signals of greater attenuation to the right. With these weak signals, the bias of 17 volts gives greatest visibility and persistence. The dimmer signals are not seen at all with the 20-volt bias.

Bias for Optimum Visibility

Figure 6 is plotted to show the visibility of signals as a function of CRT bias for pips at 5 different times of decay. It is evident that the bias for optimum visibility depends on the age of the image. For images less than 7 seconds old, the most favorable bias is 17 volts. For signals more than 7 seconds old, the bias for maximum visibility is 20 volts. At 7 seconds of decay, biases of 17 and 20 volts give about equal visibility.

Comparison of Visibility at Fluorescence and at Phosphorescence

Figure 7 is a plot showing the visual thresholds at short times of delay. This graph also includes the visual thresholds at fluorescence (a delay time of 0) which could not be depicted on the logarithmic time scales of the previous graphs. It can be seen that visibility at any CRT bias is at a maximum at either one of two intervals: either it is maximum at fluorescence or else it is greatest at the very short delay of 0.04 second. The dim scope (bias of 20 volts) gives greatest visibility at fluorescence, whereas the bright scopes (biases of 11 and 14 volts) have maximum visibility im-

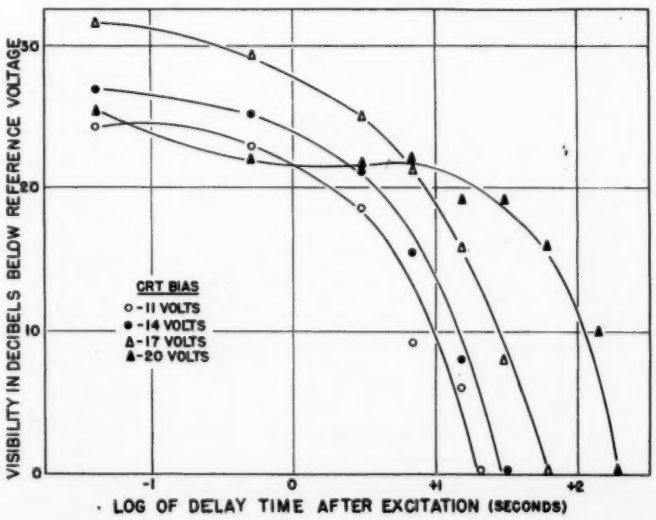


FIG. 4. Visibility of signals on a P7 screen as a function of time after excitation.

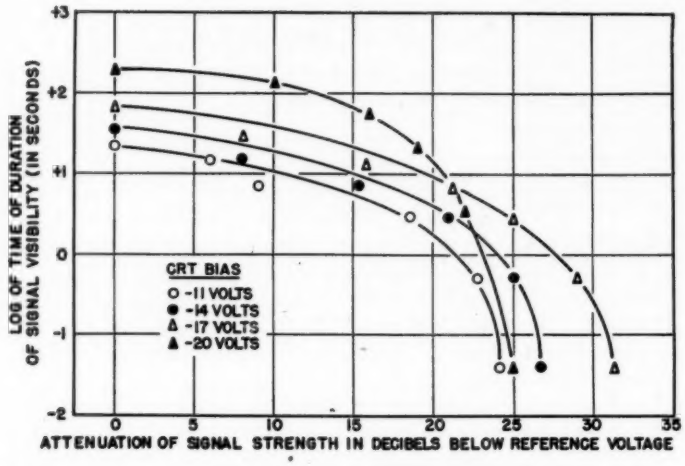


FIG. 5. Duration of visibility for signals of different strengths at four different CRT biases.

mediately following fluorescence. The bias of 17 volts is intermediate between these as there is scarcely any difference in visibility between fluorescence and the succeeding phosphorescence.

Figure 8 brings out these relationships more exactly. The ordinate represents the difference between the visual thresholds obtained at fluorescence and immediately afterwards. The graph shows that at the low biases the visibility for signals seen just behind the sweep line is greater by 4 db than when seen directly on the sweep. With the high bias (20 volts), fluorescent visibility, on the other hand, is greater by 4 db than visibility at a phosphorescence immediately ensuing. At a bias of 17 volts the difference is close to 0, showing little or no change in visibility between these two conditions.

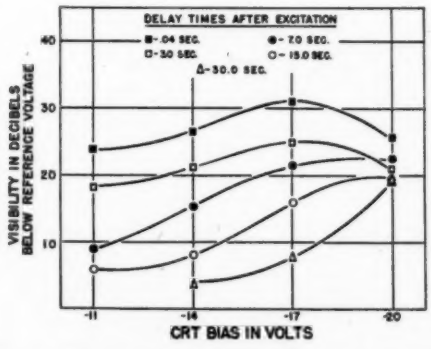


FIG. 6. Strength of minimum detectable signal as a function of different CRT biases for various times of delay.

DISCUSSION OF RESULTS

Limitation of Results to the P7 Screen

The data here reported pertain only to P7 phosphor screens and different results would be anticipated for screens with different decay characteristics. However, this limitation is not too narrow since the P7 screen is used widely on PPI scopes.

Duration of Visibility

It was found that the duration of visibility for signals depended on the CRT bias and particularly that the longer persistences occurred at the higher biases. This may be explained by taking into account two factors: (1) two separate decay functions are going on—the decay of the brightness of the background and the decay of the brightness of the signal image; (2) the visibility of the signal is in part a function of the ratio of image brightness to the brightness of the background. Now, the low CRT biases have bright scope backgrounds so that the decay of the image would be comparatively brief before the signal merged with the background. On the other hand, a high CRT bias furnishes a scope background of low luminosity so that the signal can decay over a longer interval before its contrast with the background diminishes below threshold.

That such is the case is supported by the brightness curves for the various biases (Figs. 2 and 3). Figure 2 shows that a given increment in

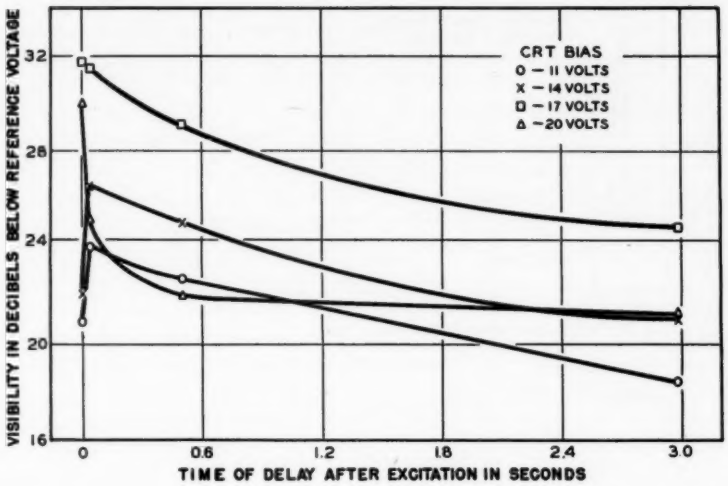


FIG. 7. Visibility of signal at fluorescence and at short intervals of phosphorescence with four different CRT biases.

signal voltage at the high biases results in a greater gain in brightness than at the low biases where the brightness curve levels off. This results in a greater contrast between signal image and scope background at the high biases which is reflected in a longer duration of visibility. Figure 9 graphically demonstrates this occurrence. Shown there is a 2-volt signal impressed on an (*a*) 11-volt bias and (*b*) 18-volt bias. It can be seen from this figure that the same voltage signal results in a greater duration for relative contrast at a high CRT bias than at a low bias. This would account for the empirical observation that there is greater persistence of visibility with high biases.

The persistence of visibility with a dark scope (bias of 20 volts) is quite marked. Strong signals (of the order of 10 db or less in this experiment), for example, were visible for more than 100 seconds. This feature might have some advantages under some circumstances, but it would be most undesirable if many objects are to be detected and if there is relative motion between the objects and the radar. The old images would then clutter the scope and interfere with the detection of most recent pips. An additional disadvantage of operating the PPI scope at a high bias is that weak signals are not seen. Preferable, in general, would be a slightly lower bias which gives a shorter persistence and which has the important advantage of permitting the greatest visibility for weak signals of recent origin.

Delay and Optimum Visibility

It was also found that for a dim scope best visibility for signals occurs at the instant the sweep lays down the image. On the other hand, with bright scopes, maximum visibility occurred when the observer looked immediately behind the sweep. Again this can be explained in terms of contrast of signal image to background. At the low CRT biases, the sweep is very bright; according to the curve relating sweep brightness to voltage (Fig. 2), a very strong signal would be required at this bias to produce appreciable contrast. Once the sweep has gone by, however, a weaker signal can be seen; though the background intensity is much less, the relative con-

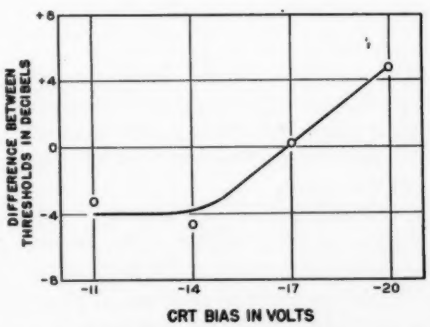


FIG. 8. Difference in visibility of signal image at fluorescence and immediately afterwards (0.04 second) (visibility at fluorescence minus visibility immediately following).

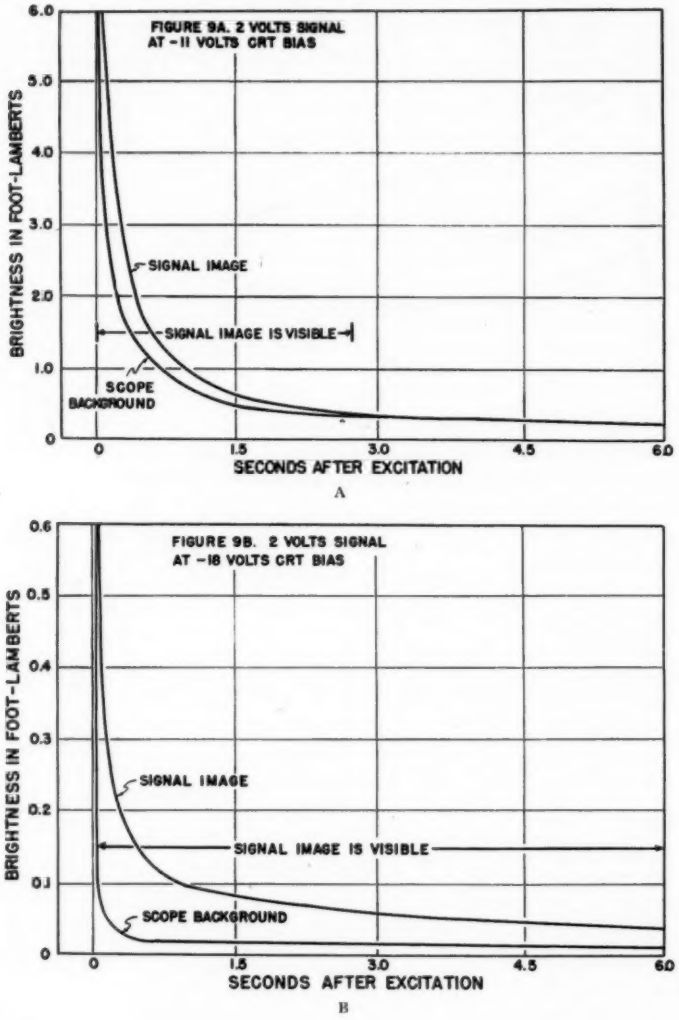


FIG. 9. Explanation of duration of visibility in terms of brightness contrast between signal image and scope background (data from brightness decay curves).

trast has actually increased. Moreover, visibility of a signal image on a bright sweep is made difficult by a glare effect which the observers frequently reported as interfering with the discernment of the signal.

The fact that at the high bias of 20 volts, the signal image is most visible on the sweep itself can be explained in terms of the eye's differential sensitivity to brightness contrast. With the dim sweep of 20-volt bias, we are operating at a brightness level (0.5 foot-lambert for the brief instant of excitation) where differential sensitivity is definitely impaired by the low intensity of brightness.[4] After the sweep passes by, the brightness level drops radically, making contrast sensitivity worse. Consequently, the better visibility is at fluorescence where the brightness level of the background is highest.

SUMMARY AND CONCLUSIONS

This experiment was designed to measure changes in visibility of a signal appearing on a P7

[4] H. R. Blackwell, J. Opt. Soc. Am. **36**, 624 (1946).

screen of a PPI scope when it is seen at different times after excitation. Different times of delay between the excitation of the screen image and its exposure to an observer were obtained by a special apparatus which could vary the angle between the rotating sweep, which excites the image, and a revolving slit which subsequently exposed it. Visibility was measured under four different scope brightnesses which were regulated by adjusting the CRT bias of the PPI tube. Visual thresholds for each of five observers were determined by variation of the strength of the signal until the subject reported a just visible signal image.

The results show that:

(1) Maximum visibility of a signal during the first seven seconds after excitation is obtained with a CRT bias such that the sweep is fairly dim yet clearly detectable. For images having decayed more than seven seconds, best visibility occurs at a higher bias furnishing a less bright sweep.

(2) Signal visibility, in general, decreases with time. The decay in visibility is most pronounced at the low biases (bright scopes) and is more gradual at high biases (dim scopes).

(3) Visibility is at a maximum for dim scopes when the signal is seen directly on the sweep. If the scope is bright, the signal is most visible just immediately behind the sweep.

(4) As the strength of the signal increases, the duration of the visibility of the signal increases. The amount of visual persistence depends on the bias and initial signal strength. For very strong signals of the order of 8 volts, the duration of visibility varied from 21 seconds (at a bias of 11 volts) to 196 seconds (at a bias of 20 volts).

International Commission of Optics

Plans are under way for the formation of an International Commission of Optics, which will be associated with the International Union of Pure and Applied Physics (I.U.P.A.P.). A preliminary meeting was held at Prague in June, 1947, and a second meeting will be held at Delft, Holland, from July 12 to 18, 1948.

It is expected to obtain financial aid from U.N.E.S.C.O. for two representatives from the United States to attend the meeting. Any member of the Optical Society who would like to attend should communicate at once with:

R. KINGSLAKE
Chairman, U. S. National Committee, I.C.O.
Hawk-Eye Works,
Eastman Kodak Company,
Rochester 4, New York

Phase Microscopy with Vertical Illumination*

HELEN JUPNIK, HAROLD OSTERBERG, AND G. E. PRIDE
Research Division, American Optical Company, Scientific Instrument Division, Buffalo 15, New York

(Received October 14, 1947)

Phase microscopy has been adapted for observing specularly reflecting specimens with the vertical illuminator. The specimen is illuminated by light, preferably monochromatic, issuing from an annular opening in a diaphragm placed at the entrance pupil of an optical system consisting of the field lens, the microscope objective, and the reflecting surface of the specimen. The diaphragm is imaged on a diffraction plate placed at the exit pupil of this optical system. Stray light and illumination losses are minimized by locating the exit pupil between the beam splitter and the eyepiece. The method can be used advantageously to examine those surface irregularities which give rise to small differences in optical path or in phase between the reflected-ray bundles. Both bright and dark contrast with excellent definition may be obtained by the use of suitable diffraction plates. The advantage of this method disappears when the specimen diffuses the incident light.

I. THE SYSTEM

THE principles of phase microscopy have been adapted to the vertical illuminator, as illustrated in Fig. 1, for viewing surfaces with good specular reflectivity.

The light source is imaged by a condenser at a diaphragm, D, which serves as entrance pupil. An image, D_1, of the diaphragm is formed beyond the specularly reflecting surface by the field lenses and the objective. The real image, D_2, formed by repassage of the light through the objective becomes the exit pupil and, therefore, the location of the diffraction plate. The planes of the diffraction plate and the diaphragm D are conjugate, as in the usual system of phase microscopy, when the object is brought into sharp focus in the plane of the eyepiece. In this manner the basic conditions[1] for obtaining phase microscopy are satisfied provided that the object does not scatter. In the event of excessive scattering, no image of the diaphragm D will be formed at the exit pupil. Relatively small amounts of scattering produce deleterious effects upon phase contrast phenomena.

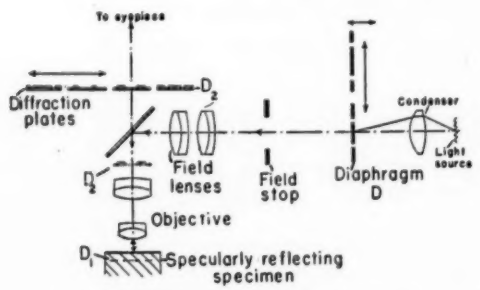


FIG. 1. Optical system of phase contrast vertical illuminator.

It is advantageous, whenever possible, to place the diffraction plate on the eyepiece side of the beam splitter. When the diffraction plate is placed on the objective side of the beam splitter, not only is the illuminating beam weakened by double passage through the diffraction plate, but also non-image forming light is reflected by the diffraction plate toward the eyepiece. These two effects become marked with the A-type[2] diffraction plate for which the absorbing zone has low transmission. Furthermore, it is mechanically convenient to locate D_2 on the eyepiece side of the beam splitter.

The area occupied in the plane of the exit pupil, D_2, by the image of the opening in the diaphragm D is called the *conjugate area* of the diffraction plate; the remaining area is called the *complementary area*. The conjugate area is of annular shape in these experiments. The inner and outer radii of the annular conjugate area, and, correspondingly, of the opening in the diaphragm D, are chosen different for objectives of different focal length. These diaphragms are

* Presented at the Thirty-First Annual Meeting of the Optical Society of America, New York, New York, October 4, 1946.

[1] Bennett, Jupnik, Osterberg, and Richards, Trans. Am. Micro. Soc. LXV, 101, paragraph 1. (1946).
[2] See reference 1, p. 106.

centerably mounted in a slide for interchangeability when shifting objectives. In addition, the diaphragm slide is adjustable in the direction of the optical axis for the purpose of focusing D upon a selected position for D_2. The slide D also contains a circular opening to allow the vertical illuminator to be used with no diffraction plate but with standard, central-cone illumination. The circular opening is an iris which acts as the aperture stop.

Diffraction plates are mounted in similar slides, each of which carries four centerable diffraction plates and one glass blank. In this way the diffraction plates can be rapidly interchanged in the comparison of images. The two sliding carriers are provided with positioning stops for each diffraction plate or diaphragm.

A series of diffraction plates which differ in light transmission and in optical path in the conjugate and complementary areas is provided for an objective of 16-mm focal length, N.A. 0.25, and for an objective of 4-mm focal length, N.A. 0.85. Since previous experience in phase microscopy has shown that the conjugate area of many of the useful diffraction plates has low light transmission, the diffraction plate and exit pupil, D_2, are located on the eyepiece side of the beam splitter. Preliminary examination of the available object specimens verified that the eyepiece side was the preferred location. In these experiments vignetting of the field by the condenser diaphragm caused little difficulty whether D_2 was on the eyepiece or objective side of the beam splitter.

The field stop is focusable, as in the standard vertical illuminator, but the diaphragms D with annular opening serve as fixed rather than variable aperture stop.

A special vertical illuminator was constructed to contain all of the above mentioned elements with the exception of the source of light and condenser. Either a tungsten filament or an H-4 mercury arc filtered for the 5461A line is used as source of light. Best phase contrast is obtained with the H-4 source, although at the expense of reduced intensity and increased optical difficulties.

II. METHOD OF ADJUSTMENT

The specularly reflecting specimen is mounted upon the stage of the microscope and is made mechanically perpendicular to the optical axis of the microscope. The specimen is brought into focus after the illuminating system has been adjusted. Assume that the diaphragm D and the diffraction plates require adjustment. The eyepiece is then replaced by a short-focus telescope which is focused upon the diffraction plate. The field of this telescope is greater than the clear aperture of the objectives and so permits centering of the diffraction plates with respect to the clear aperture. The four diffraction plates in the slide are individually centered by means of centering screws. The image of the annular opening of the corresponding diaphragm D is next focused upon the conjugate area of one of the centered diffraction plates by means of the lateral and longitudinal adjustments of the diaphragm. If an objective of given focal length is provided with two sets of diffraction plates having annular conjugate areas of different size, a diaphragm D is required for each set. These adjustments of the diffraction plates and diaphragms do not require repetition but should be checked occasionally. The eyepiece is replaced, the light beam further adjusted for uniform illumination, and the specimen observed. A change in specimens necessitates reorientation of the surface perpendicular to the optical axis in order to avoid lateral displacement of the image of the diaphragm D.

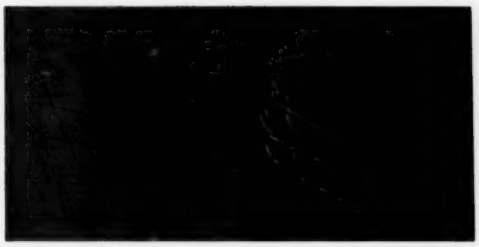


FIG. 2. Scratches in aluminum first surface mirror. Magnification 90×. A = no phase contrast. B = bright contrast. 0.12 A+0.25λ diffraction plate.

III. OBSERVATIONS

It has been found that the phase-vertical illuminator can increase contrast in the image of surface details which give rise to small differences in optical path between rays reflected from neighboring points on the specimen. In

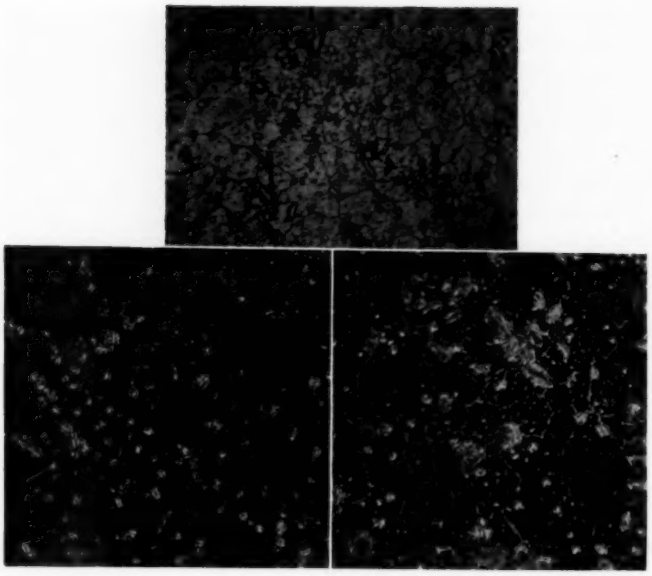


FIG. 3. Grain boundaries in high speed steel. Magnification 500×. *A* = no phase contrast. *B* = dark contrast. 0.12A −0.25λ diffraction plate. *C* = bright contrast. 0.12A+0.25λ diffraction plate. Elevated carbide segregations are also visible.

such cases either bright or dark contrast can be obtained by suitable choice of diffraction plate. As is to be expected, the method cannot be used to advantage to examine diffusing surfaces or those details which are easily made visible by the use of standard methods because of differences in reflectivity or polarization properties. Examples of surface details for which the method is useful are shallow surface irregularities such as fine scratches and sleeks, neighboring areas which display small differences in phase on reflection, neighboring areas which give rise to small optical-path differences resulting from penetration and subsequent reflection of the incident light, and neighboring areas which show small differences in reflectivity or in absorption either with or without a small change in phase. The method may prove to be most useful for examining well polished surfaces of isotropic or weakly anisotropic materials for which the polarizing vertical illuminator is not successful.

Some of the results obtained are shown in the following figures. The use of conical illumination will be mentioned in connection with some of the photo-micrographs taken without a diffraction plate. In these cases, the conical illumination was obtained simply by leaving the annular diaphragm *D* in the illuminating beam. The use of conical illumination frequently resulted in a specimen's image showing greater contrast than that seen when the observation was made by using the full aperture of the objective. The imagery and increase in contrast so obtained was judged to be at least equivalent to that obtained with central-cone illumination, but with the aperture suitably reduced.

Fine scratches and sleeks in a specularly reflecting surface serve as an excellent test object for determining whether phase microscopy is possible with vertical illumination and, if so, whether it agrees with theoretical expectations. Figures 2a and b illustrate superficial scratches in an aluminum mirror. In Fig. 2a the objective of 16-mm focal length with no diffraction plate was used together with central-cone illumination and reduced aperture. Many deep scratches are visible, but finer scratches which are not to be identified positively are present. Figure 2b shows the same field of view taken with a 0.12A+λ/4 diffraction plate in position; i.e., with a diffraction plate for which the transmission ratio of the conjugate area to the complementary area is 12 percent and for which the light passing through the conjugate area is retarded in phase by one-quarter wave-length with respect to the light passing through the complementary area.

As expected,[3] the scratches are seen in bright contrast. The finer scratches and sleeks are much more visible than in Fig. 2a taken with no diffraction plate. A $0.12A-\lambda/4$ diffraction plate was also used. This diffraction plate differs from the $0.12A+\lambda/4$ plate in that the light passing through the conjugate area is advanced in phase by one-quarter wave-length with respect to the light passing through the complementary area. The scratches appeared in dark contrast, as expected,[3] and the finer detail was much more visible than without the use of a diffraction plate. In the visual observations comparisons are made at the position of sharpest focus. Although reversal of contrast can be effected by incorrect focusing, this procedure invariably reduces contrast.

A similar example is provided by the grain boundaries present in a section of high speed steel. Aqua regia in methanol was used to reveal the austenitic grain size. Figure 3a is a photograph of the specimen taken with the use of an objective of 4-mm focal length and conical illumination. To obtain Fig. 3b the $0.12A-0.25\lambda$ diffraction plate was inserted, and to obtain Fig. 3c the $0.12A+0.25\lambda$ diffraction plate was inserted. Figures 3b and c show these grain boundaries in dark and bright contrast, respectively, according to expectation.

The objective of 4-mm focal length was used to obtain Figs. 4a and b which show surface variations resulting from stain on a prism of dense flint glass. Figure 4a was taken with no diffraction plate but with conical illumination. The stained area contains groups of very small spots which appear to be arranged in layers. The uppermost layer of black spots is easily seen, but the boundaries of the two lower layers are not well defined, especially in the case of the two smaller groups. These boundaries are better defined with a diffraction plate which causes the spots to appear in bright contrast, as in Fig. 4b which was obtained with a $0.12A-\lambda/4$ diffraction plate. With this diffraction plate the scratches appear dark as expected. The bright spots are areas on which the optical path of

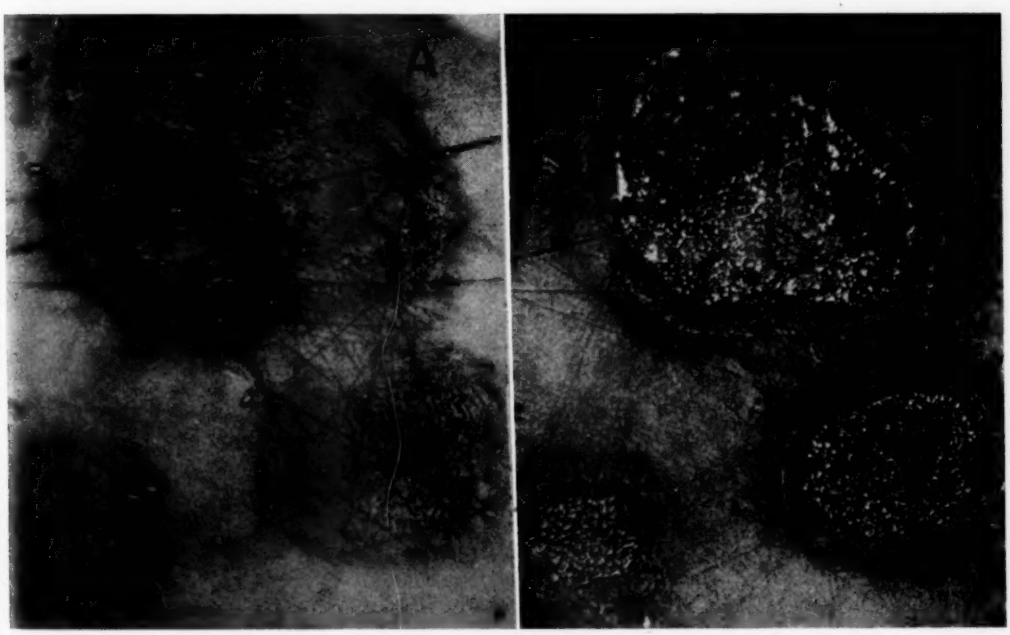


FIG. 4. Stains on dense flint prism. Magnification 300×. *A* = no phase contrast. *B* = bright contrast. $0.12A-0.25\lambda$ diffraction plate.

[3] See reference 1, p. 117.

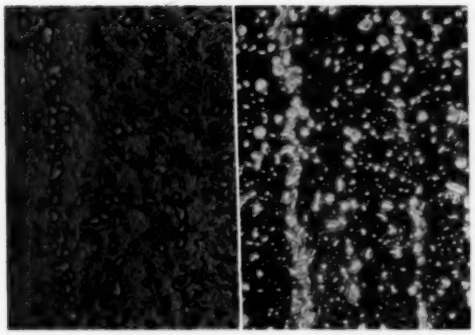


FIG. 5. Spheroidized carbides in high carbon steel. Magnification 500×. A = no phase contrast. B = bright contrast. 0.12A − 0.25λ diffraction plate.

the reflected light has been decreased with respect to that of the light reflected from the surrounding darker areas. The assumption that the bright spots are elevations on a leached surface is at least consistent with this conclusion. Both the scratches and the bright spots of Fig. 4b were strongly reversed in contrast with the use of a 0.12A + λ/4 diffraction plate.

Figures 5a and 5b show carbide segregations in a high carbon steel. Figure 5a was taken with an objective of 4-mm focal length and central cone illumination with reduced aperture. Figure 5b is a photograph of the same region of the specimen when observations are made with a 0.12A − 0.25λ diffraction plate in the objective of 4-mm focal length. The microscope was focused on the smallest bright particles. The bright structures are spheroidized carbides which remained as elevations after etching with nitric acid in alcohol. Another such example has been reported by F. W. Cuckow[4] since the appearance of an abstract[5] by the writers. The elevated surfaces of martensite fragments are seen in bright contrast against the more deeply etched "intermediate transformation product".

Mention is made of the fact that the particular diffraction plates used in these experiments resulted in some loss of depth of focus. This is apparent, for example, in the photographs of Figs. 3 and 4. Also, in the case of the observed metallurgical specimens, complete reversal of contrast by shifting from an $A+$ type plate to an $A-$ type plate was possible only for smaller details. It is interesting, for example, to note that with the available diffraction plates it was not possible to reverse the bright contrast of the larger carbide particles seen in Figs. 3 and 5.

IV. CONCLUSIONS

Only a preliminary and partial investigation of the usefulness of the phase-vertical illuminator has been made, but the experiments indicate that the range of applicability is in agreement with theoretical expectation. The use of this method is restricted by the requirement that the polished or etched surface shall have little scattering.

[4] F. W. Cuckow, Nature **159**, 639 (1947).
[5] H. Jupnik, H. Osterberg, and G. E. Pride, J. Opt. Soc. Am. **36**, 710A (1946).

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 38, NUMBER 4 APRIL, 1948

Centering of Optical Systems

A. L. INGALLS AND K. PESTRECOV
Bausch & Lomb Optical Company, Rochester, New York

(Received August 11, 1947)

An autocollimation method has been developed whereby the centering of component surfaces within an optical system may be measured. Paraxial transfer formulas containing a quantity called eccentricity are developed after definitions have been promulgated for certain terms. For a given optical system, the position, with respect to a reference axis, of an image due to reflection from an optical surface within the system depends on (*a*) the distance of the object from the reference axis, (*b*) the magnification of the reflecting and refracting surfaces involved and (*c*) the eccentricities of each surface. If all other factors are known, the eccentricity of each surface may be computed from measurements of the displacement of the autocollimated image.

With equipment commonly available certain optical systems may be examined and with proper instrumentation other systems, including bent systems utilizing reflecting prisms or mirrors, may be inspected. A few possible applications are as follows:

1. adjusting a lens on an edging machine;
2. measuring the eccentricity of lens surfaces with its edge;
3. measuring lens eccentricity on a lathe during mounting;
4. measuring the eccentricities of the various surfaces of a photographic lens;
5. checking the centering of a telescope, binocular or periscope;
6. decentering an optical element by a known amount in order to determine the effect of eccentricity on the performance of the system.

INTRODUCTION

IN order to center an optical system such as a photographic lens which has several air-glass surfaces, one has first to define what he means by centering. The generally accepted theoretical concept is that an optical system is centered if the vertex centers of curvatures all lie on one straight line, and all the plane surfaces contained in the system are perpendicular to this line which is called the axis. This definition applies, of course, also to "bent" systems in which one common axis is realized by reflection at inclined surfaces (prism, mirrors).

However, the perfect condition of centering in lens making rarely occurs any more frequently than perfection in any other manufactured product. Tolerances are provided on such things as edging, cementing, boring, turning, etc., any of which may be eccentric as well as off diameter. This presents the problems of mechanical eccentricity and loose fits.

In the presence of eccentricity, the axis as defined above does not exist. Consequently, if tolerances are to be established for centering, a new definition of an axis must be made before the tolerance of eccentricity of any element or surface may be specified or the amount measured in a lens. Such a straight line must be defined in order to be useful. Here is a point of difference between the designer, the manufacturer, and the user. The designer assumes a line in space and centers his surfaces about that line. He then specifies how much a lens element or surface may be decentered from that line. The manufacturer assumes an axis which is the axis of rotation of his machine, such as an edging machine or a lathe and locates his lens elements on that axis. The user, however, uses an axis dependent on the mounting thread and shoulder of the finished lens. If the shoulder is made perpendicular to the thread axis, the axis is positive. If, however, the shoulder is not perpendicular to the thread axis, the position of the axis is more difficult to determine. If the camera thread is tight, the thread axis is the more important. If, however, the thread is loose, the lens usually sits tightly against its shoulder. The axis then becomes a line through a point centered with the thread and perpendicular to the plane of the mounting shoulder.

After a lens is made, it is difficult to use the designer's axis. In actual inspection, it would be necessary first to assume any logical axis, as, for instance, a line passing through the centers of curvature of the first and second, or the first and last surfaces. Next, the surfaces would have to be measured to find their eccentricities to

that axis, and then a new axis chosen such that all or the greatest number of surfaces fall inside of their eccentricity tolerances. It then remains necessary to determine how closely this axis comes to being coincident with the axis formed by the mounting thread and shoulder of the lens.

The manufacturer's axis of rotation is a good axis but may not agree with the designer's or the user's axis, depending on the errors he makes in centering the lens elements and how true the mounting thread and shoulder conform to his axis of rotation.

Therefore, it seems more convenient to define "the axis of the complete lens" as "the axis of the final mounting thread." This is a good axis to choose, from a practical standpoint, because it is much easier to true the shoulder to the thread than to true the thread to the shoulder. As a matter of fact, the shoulder and thread should be made at the same lathe-chucking to insure that they be true. If later the shoulder is recut to set the shoulder to focal plane distance (parfocalize), much care should be exercised to keep the shoulder plane perpendicular to the thread axis.

Since the designer assigns tolerances and the manufacturer must make a lens so that it is a satisfactory lens to the user, they must adhere to this same axis.

Definitions.—The lens axis of a completely finished photographic lens shall be its final mechanical mounting axis. It will be known here as the reference axis. Having thus defined the axis of a mounted compound lens, the eccentricity of its optical surfaces and lens elements may be defined.

Surface axis is defined as a line parallel to the reference axis and passing through the vertex center of curvature of the surface.

An eccentric lens surface is one in which the vertex center of curvature lies off the reference axis by an amount called the eccentricity. It is the distance from the center of curvature to the reference axis measured along the normal to the reference axis.

An eccentric lens element is one in which the two vertex centers of curvature of its two surfaces lie equal distances in the same direction from the reference axis.

A tipped or tilted lens element is a lens element in which the two centers of curvature lie at unequal distances or directions from the reference axis.

CENTERING METHODS

Being now equipped with definitions, one may talk about centering methods. There are three methods of centering lenses in general use today as follows: (*a*) mechanical, (*b*) transmitted beam, and (*c*) reflected beam.

The mechanical method consists of clamping the lens element between two concentric tubes and depending on the lens to center itself between the tubes.

The transmitted beam method makes use of an object and an image formed by the refraction of the lens element. The element is rotated so that any decentration will cause the image to move in a circular pattern. This method is defective in that the lens may be tipped but still produce no deviation in the image, as long as the transverse axis of the lens coincides with the axis of rotation.

The reflected beam method is probably the most sensitive and widely used but has certain difficulties which may be remedied as suggested in this paper. Looking into a lens, with or without the aid of a magnifier, at the reflected images of the window bars or another distant light source is the method commonly used. When the lens is rotated, any eccentric surfaces will be indicated by a circular motion of one or more of the images. The difficulty is that the motion of the images from inside surface reflections are influenced by the eccentricities of the outer surfaces. Thus, the image from a perfectly concentric inside surface may move because an outside surface is eccentric. Or the inside surface may be eccentric, but the reflected image may stand still because the eccentricity of an outer surface compensates for it. Another uncertainty arises from the fact that it is difficult to tell the amount of eccentricity from the motion of the image of a distant source. For example, a plane surface will appear to have a large eccentricity for small errors whereas a highly curved surface will appear to have very little eccentricity when the error is actually large.

QUANTITATIVE REFLECTION METHOD

It has been found possible to measure the eccentricities of the surfaces of a completed lens with sufficient accuracy either to secure a desired centering in manufacturing or to compare eccentricities of finished lenses with blue print tolerances.

The theoretical basis of this method may be outlined as follows. Referring to Fig. 1, O is an object (light source) located in the vicinity of the reference axis, O' is its image formed by a refracting (or reflecting) surface whose center of curvature C is at distance e from the reference axis (it is positive if C is above this axis). The distance of object O from the reference axis is h, and the distance of image O' is h' (they are also positive if above the axis). Since in well mounted objectives e is always very small and since h is purposely set small by a proper instrumentation (to be described later), the following paraxial relationship holds:

$$(h'-e)/(h-e)=m,$$

where m is the image magnification.

For refracting surfaces, $m=nl'/n'l$; and for reflecting surfaces, $m=-l'/l$.

Solving the above equation for h', we obtain:

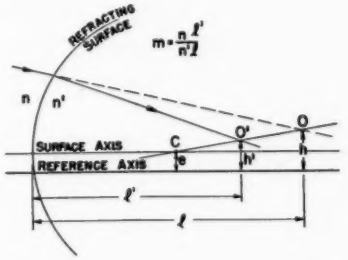
$$h'=(1-m)e+mh \qquad (1)$$

This equation may be applied to a number (k) of successive surfaces. Thus we obtain:

$$\begin{aligned}
h_1' &= (1-m_1)e_1+m_1h_1\\
h_2' &= (1-m_2)e_2+m_2h_1'\\
&= (1-m_2)e_2+m_2(1-m_1)e_1+m_2m_1h_1\\
h_3' &= (1-m_3)e_3+m_3h_2'\\
&= (1-m_3)e_3+m_3(1-m_2)e_2\\
&\quad +m_3m_2(1-m_1)e_1+m_3m_2m_1h_1\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
h_k' &= (1-m_k)e_k+m_k(1-m_{k-1})e_{k-1}\\
&\quad +m_km_{k-1}(1-m_{k-2})e_{k-2}+\cdots\\
&\quad +(m_km_{k-1}\cdots m_2)(1-m_1)e_1\\
&\quad +(m_km_{k-1}\cdots m_1)h_1. \qquad (2)
\end{aligned}$$

This general form indicates that, for k surfaces, there will be $k+1$ terms on the right side of the equation.

Assuming the original propagation of light from left to right, the image formation, utilized in our method of measuring the eccentricity of the kth surface, will involve the passage of light through $k-1$ surfaces from left to right, then reflection from the kth surface, and finally the passage of light from right to left again through $k-1$ surfaces. Thus, effectively $2k-1$ surfaces will take part in the image formation. Consequently, as is indicated by Eq. (2), the expression for the final h' (which is intended to be measured and then used for the eccentricity computations) should contain $2k$ terms. However, a considerable simplification is possible if the instrumentation is set up so that the final image is produced in the plane of the original object (light source). Under this condition of autocollimation, the over-all magnification becomes equal to minus one (-1), and the magnification of the reflecting surface k, whose eccentricity is to be measured, necessarily becomes also minus one (-1). Furthermore, under this condition, the magnification of an image formed by any surface with the light direction from right to left is the reciprocal of the magnification produced by the surface the first time, i.e., with the left to right imagery. Denoting $1/m_k=r_k$ and double-priming the distances h, produced in the right to left imagery



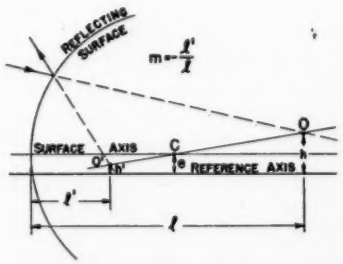


FIG. 1a. A refracting surface eccentric with respect to the reference axis. C is its center of curvature. O is an object point imaged paraxially at O'.

FIG. 1b. A reflecting surface eccentric with respect to the reference axis. C is its center of curvature. O is an object point imaged paraxially at O'.

after the reflection at the kth surface, we obtain in accordance with Eq. (1):

$$h_{k-1}'' = (1-r_{k-1})e_{k-1} + r_{k-1}h_k'$$
$$h_{k-2}'' = (1-r_{k-2})e_{k-2} + r_{k-2}h_{k-1}''$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$h_1'' = (1-r_1)e_1 + r_1h_2''.$$

By using the value of h_k' as given by Eq. (2), and remembering that in this set up (surface k is autocollimating) $m_k=-1$ and for any other surface $1/m=r$, we obtain after a successive elimination of the intermediate (h'') terms and after collecting similar terms:

$$h_1'' = 2(1-r_1)e_1 + 2r_1(1-r_2)e_2 + 2r_1r_2(1-r_3)e_3 + \cdots + 2(r_1r_2r_3\cdots r_{k-2})(1-r_{k-1})e_{k-1} + 2(r_1r_2r_3\cdots r_{k-1})e_k - h_1. \quad (3)$$

Now we have only $k+1$ terms in the expression determining the final $h'=h_1''$ obtained after a double passage through $k-1$ surfaces and a reflection at the kth surface.

Equation (3) may be written in the following form:

$$h' = K_1e_1 + K_2e_2 + K_3e_3 + \cdots + K_{k-1}e_{k-1} + C_ke_k - h, \quad (4)$$

where the general coefficient $K_{k-1}=2(r_1r_2r_3\cdots r_{k-2})(1-r_{k-1})$ and the coefficient for the term containing e_k is:

$$C_k = 2(r_1r_2r_3\cdots r_{k-1}).$$

In order to use properly the general form given by Eq. (4), it should be clearly understood that eccentricities of all the surfaces from the first to the $(k-1)$ are associated with respective coefficients K, while the term containing the eccentricity of the reflecting (i.e., autocollimating) surface always has coefficient C_k. Furthermore, it should be noted that when the first surface is autocollimating ($k=1$), the correct evaluation of the coefficients is: $K_1=0$ and $C_1=2$; when the second surface is autocollimating ($k=2$), the correct evaluation is: $K_1=2(1-r_1)$, $K_2=0$, $C_2=2r_1$. For any other number (k) surfaces with the kth surface autocollimating, the evaluation of the first coefficient remains $K_1=2(1-r_1)$, and the evaluation of all other coefficients K is obtained by a successive substitution of k, in the expression for K_{k-1}, by all the integers from 3 to k.

For example, for four surfaces with the fourth surface autocollimating, the coefficients evaluate as follows:

$$K_1 = 2(1-r_1),$$
$$K_2 = 2r_1(1-r_2),$$
$$K_3 = 2r_1r_2(1-r_3),$$
$$C_4 = 2r_1r_2r_3.$$

Finally, it should be emphasized, that all the coefficients (K and C) are not intrinsic constants for a given lens formula. They acquire different values depending on the surface utilized for autocollimation.

Equation (4) shows that in order to find the eccentricity of the kth surface, the eccentricities of all the preceding surfaces should be known. This indicates that the actual measuring procedure should consist of k collimations starting with the first surface and ending with the kth. Thus a complete analysis of a lens involves the computation of K and C coefficients for each surface in each collimation, the measurement for each collimation of the final image distance (h') from the reference axis, and the successive evaluation of eccentricities (e_k) from the measured values h' and the computed coefficients K and C.

The evaluation may be somewhat simplified by placing the object on the reference axis, as then h becomes zero and drops out of Eq. (4). Under this condition (actually utilized in our instrumentation), the eccentricities are represented by the following expressions:

first surface autocollimating:
$$e_1 = h'/C_1 = h'/2,$$
second surface autocollimating:
$$e_2 = (h' - K_1e_1)/C_2,$$
third surface autocollimating:
$$e_3 = (h' - K_1e_1 - K_2e_2)/C_3,$$
etc.

To avoid a misunderstanding, it should be repeated here that h', representing the final image distance from the reference axis, will generally acquire different values as we proceed from surface to surface, and that the coefficients K will have different values for each autocollimation.

A plane refracting or reflecting surface presents a special problem. Since a displacement of a plane surface in a direction perpendicular to the reference axis has no effect on performance, the only case of significance is that of a tipped surface. Since the radius of curvature of a plane surface is infinity, the eccentricity (as previously defined) of a tipped plane surface is always also infinity ($e=\infty$). If we substitute this value of e into the formulas derived above, we find that, for a plane refracting surface the expressions containing $(1-m)e$ and $(1-r)e$ become indeterminate because the lateral magnification (m) of a plane refracting surface is equal to one, and, consequently, the products under consideration become zero times infinity.

If the plane surface (k) is autocollimating, its object is at infinity. Hence, the preceding surface ($k-1$) focuses at infinity; its magnification (m_{k-1}) is also infinity and its $r_{k-1}=0$. Consequently, the expression $2(r_1r_2r_3\cdots r_{k-1})e_k$ of Eq. (3) becomes indeterminate.

The following paraxial analysis is used to evaluate these indeterminate forms.

In Fig. 2, the plane refracting surface is tipped by angle δ from the normal to the reference axis. This angle is considered positive if the plane is tipped counter-clockwise.

An element of a plane surface may be considered as an element of a spherical surface of an arbitrary large radius "R," then: $e=R\delta$. Now Eq. (1) may be rearranged by making the above substitution and also substituting $m=nl'/n'l$.

$$\begin{aligned} h' &= (1-m)e+mh \\ &= (1/m-1)me+mh \\ &= (n'l/nl'-1)Rm\delta+mh. \end{aligned}$$

But from $n'/l'=n/l+(n'-n)/R$ we find by rearrangement that $(n'l/nl'-1)R=(n'/n-1)l$.

Therefore:

$$h'=(n'/n-1)ml\delta+mh. \qquad (5)$$

If we designate $E=-l\delta$ and $n'/n=M$ (which is the longitudinal magnification produced by a plane refracting surface), and considering the fact that for a plane refracting surface the lateral magnification $m=1$, the above equation may be rewritten in the following form:

$$h'=(1-M)E+mh. \qquad (6)$$

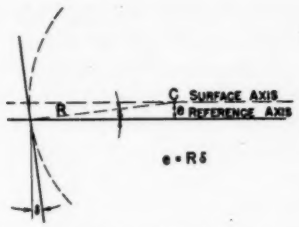


FIG. 2. A plane surface, tipped with respect to the reference axis, and a spherical surface element of an arbitrary radius R. The relationship $e=R\delta$ is used in evaluation of the indeterminate expressions for a plane refracting surface.

This transfer formula permits us to use Eqs. (1) and (2) also if one or several refracting surfaces are plane, provided that in $(1-m)e$ expressions for these surfaces we substitute M for m and E for e, and in the expressions containing products of m we use $m=1$ for plane surfaces. It may be noted that m in the last term of Eq. (6) is, obviously, also equal to one; we have, however, retained this m in order to effect a complete formal similarity of this equation with Eq. (1).

Some additional transformations are necessary in order to extend the same similarity also to the final Eq. (3), which is the working equation for actual eccentricity determinations. Indeed, in the right to left imagery, produced after the reflection from an autocollimating surface, quantity h of Eq. (5) should be replaced by h'' in accordance with the previously established notation. Then, solving this equation for h'', we obtain:

$$h''=h'/m-(n'/n-1)l\delta$$

or

$$h''=-(1-n/n')n'l\delta/n+h'/m.$$

Since, however, for plane refracting surfaces $n'l/n=l'$, and designating $-l'\delta=E'$,

$$n/n'=1/M=M',$$

and $1/m=r$ (as was previously stated, $m=1$ for plane refracting surfaces; hence, r is also equal to one), we obtain the following expression similar to that derived for curved surfaces:

$$h''=(1-M')E'+rh'. \qquad (7)$$

If we now use Eqs. (6) and (7) in the process of successive eliminations employed in the deriva-

tion of Eq. (3), we will find that this equation is also applicable when one or several refracting surfaces are plane, provided that in $(1-r)e$ expressions for these surfaces we substitute $M'=n/n'$ for r, and $E'=-l'\delta$ for e; and in the expressions containing products of r we use $r=1$ for the plane surfaces. This equally applies to Eq. (4).

The case of a tipped plane surface autocollimating presents a different problem, as it involves the imagery of the preceding surface as well. In Fig. 3, the surface k is a plane autocollimating surface tipped at an angle δ from the normal to the reference axis. An auxiliary axis is then drawn through the center of curvature of the surface $(k-1)$ perpendicular to the surface k. This auxiliary axis will thus form an angle δ with $(k-1)$ surface axis. As before, angle δ is considered positive if produced by a counterclockwise rotation. The object distance from the reference axis is h_{k-1}. The image distance h_{k-1}'' is obtained after passage of light left to right through the $(k-1)$ surface, autocollimated reflection from the kth surface and then right to left passage through the $(k-1)$ surface. The object distance l_{k-1} measured to the left from the $(k-1)$ surface along its axis is negative in the case illustrated in Fig. 3.

It should be noted that δ is considered to be a very small angle and, therefore, $\cos\delta=1$ is a valid approximation. Hence, the object plane (h_{k-1}), coinciding with the image plane (h_{k-1}''), is not only perpendicular to the auxiliary axis but, for all practical purposes, also to the reference axis. The auxiliary axis and the reference axis intersect this object-image plane in two points

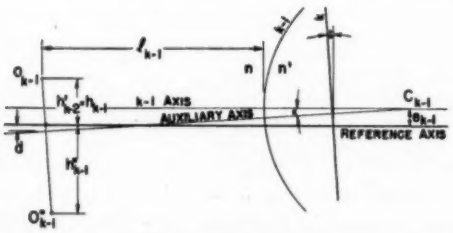


FIG. 3. A plane autocollimating surface (k) tipped with respect to the reference axis. The auxiliary axis is perpendicular to surface k. Because δ is small, $h_{k-1}h_{k-1}''$ plane is practically perpendicular to the reference axis as well as the auxiliary axis. Object point O_{k-1} is paraxially imaged as O_{k-1}'' by refraction at $k-1$, reflection at K, and second refraction at $k-1$.

separated by distance d (negative in the illustrated case).

Since $m=-1$ for an autocollimating system, we obtain from Fig. 3:

$$d-h_{k-1}''=h_{k-1}-d, \quad h_{k-1}''=2d-h_{k-1}.$$

Also from Fig. 3 we obtain :

$$\delta_k=(e_{k-1}-d)/(R_{k-1}-l_{k-1}).$$

Solving for d, it follows that:

$$d=(l_{k-1}-R_{k-1})\delta_k+e_{k-1}.$$

Substituting in the equation for h_{k-1}''

$$h_{k-1}''=2(l_{k-1}-R_{k-1})\delta_k+2e_{k-1}-h_{k-1}.$$

For this case $l'=\infty$ and, therefore, $n'/l'=0$ in the paraxial equation $n'/l'=n/l+(n'-n)/R$. Solving for R,

$$R=-(n'-n)l/n.$$

Therefore:

$$(l_{k-1}-R_{k-1})=l_{k-1}(n'/n)_{k-1}.$$

By substitution, and since $h_{k-1}=h_{k-2}'$,

$$h_{k-1}''=2(n'/n)_{k-1}\delta_k l_{k-1}+2e_{k-1}-h_{k-2}'.$$

Substituting the value of h_{k-2}' as determined by Eq. (2), we obtain:

$$\begin{aligned}h_{k-1}''=2e_{k-1}&+2(n'/n)_{k-1}\delta_k l_{k-1}-(1-m_{k-2})e_{k-2}\\&-m_{k-2}(1-m_{k-3})e_{k-3}-m_{k-2}m_{k-3}\\&\times(1-m_{k-4})e_{k-4}-\cdots-(m_{k-2}m_{k-3}\cdots\\&\times m_2)(1-m_1)e_1-(m_{k-2}m_{k-3}\cdots m_1)h_1.\end{aligned}$$

Now we may continue with a series of equations:

$$\begin{aligned}h_{k-2}''&=(1-r_{k-2})e_{k-2}+r_{k-2}h_{k-1}''\\h_{k-3}''&=(1-r_{k-3})e_{k-3}+r_{k-3}h_{k-2}''\\&\cdots\cdots\cdots\cdots\cdots\cdots\\h_1''&=(1-r_1)e_1+r_1h_2''.\end{aligned}$$

By eliminating the intermediate values of h'' and after collecting similar terms, we finally obtain:

$$\begin{aligned}h_1''=2(1-r_1)e_1&+2r_1(1-r_2)e_2+2r_1r_2(1-r_3)e_3\\&+\cdots+2(r_1r_2\cdots r_{k-2})(1-r_{k-1})e_{k-1}\\&+2(r_1r_2\cdots r_{k-2})(n'/n)_{k-1}\delta_k l_{k-1}-h_1.\end{aligned}$$

This form is exactly the same as Eq. (3) with the exception of the term next to the last, in which the product $r_{k-1}e_k$ now evaluates to $(n'/n)_{k-1}\delta_k l_{k-1}$. With this reservation, Eqs. (3) and (4) may be also used when the autocolli-

mating surface is plane. Thus, if the tip angle δ is treated as a quantity analogous to eccentricity e of curved surfaces, coefficient C_k (of the term containing the eccentricity of the autocollimating surface) in Eq. (4) acquires the following form for a plane surface (k) autocollimating:

$$C_k = 2(r_1 r_2 r_3 \cdots r_{k-2})(n'/n)_{k-1} l_{k-1},$$

while the K_{k-1} coefficients preserve the same form as for curved surfaces.

It should be noted that these evaluations cease to be applicable when a plane autocollimating surface is the front surface in a system whose eccentricities are to be measured. In this case, the tip angle δ of the front surface may be determined by the generally known autocollimation method.

When a plane autocollimating surface is the second in the system, the correct evaluation of the coefficients is $K_1=2$, $K_2=0$, $C_2=(n'/n)_1 l_1$.

Otherwise, when a plane autocollimating surface is the third or the kth ($k>3$) in the system, the evaluation of the first coefficient is $K_1 = 2(1-r_1)$, and the evaluation of all other coefficients K is obtained by a successive substitution of k, in the expression for K_{k-1}, by all the integers from 3 to k.

OPERATION OF METHOD

The use of this method is fairly simple if one has certain equipment commonly available. The purpose of such equipment is to produce an object at such a point in the optical system, whose eccentricities are to be investigated, that after reflection the image is in the same plane as the object. One of the simplest instrumentations is that actually used by us to test out the idea. This consisted of an ordinary microscope fitted with a filar micrometer, an objective with a long working distance, and a vertical illuminator. A pinhole was placed in the vertical illuminator at such a position that it was focused in the object plane of the microscope. The lens to be tested was placed on the microscope stage with its front surface toward the microscope. A scale was attached to the microscope slide to indicate the computed positions of the reflections. This quickly locates the reflection wanted among the unnecessary multiple reflections actually appearing. The deviation of each image was then measured and the corresponding eccentricity calculated from the formula and table.

Other arrangements may be used such as a microscope set coaxially with a lathe axis. A system may also be constructed with a movable lens in the object plane of a microscope and focus settings of the system obtained from plus to minus infinity by simply moving the auxiliary lens. With this equipment almost any optical system of lenses could be checked.

Since eccentricities may occur in any direction perpendicular to the reference axis, the image distance components from this axis may be measured in mutually perpendicular directions to each other and the reference direction. In Cartesian coordinates, X and Y would be used, assuming Z is the reference axis. Eccentricities x and y could then be plotted graphically to determine the position of the average lens axis with respect to the reference axis. A few possible applications are as follows:

1. adjusting a lens on an edging machine;
2. measuring the eccentricity of lens surfaces with its edge;
3. measuring lens eccentricity on a lathe during mounting;
4. measuring the eccentricities of the various surfaces of a photographic lens;
5. checking the centering of a telescope, binocular or periscope.

Conversely, another application of the method is the use of especially constructed test mounts to introduce known amounts of eccentricity and tilt into an optical system to determine the effect on performance.

A New Determination of the Young-Helmholtz Primaries

GEORGE SZEKERES
Shanghai, China

(Received July 6, 1947)

This is an attempt to determine the primaries corresponding to the fundamental sensations of the Young-Helmholtz theory. Unlike previous attempts by various authors, this determination is based solely on color-mixture data pertaining to the trichromatic observer. It is shown that the primaries can be derived from the color-mixture data of the average trichromate with reasonably small uncertainty, if only the one hypothesis is made that each of the three response functions has a certain simple geometrical character, resembling the standard luminosity function. This hypothesis is suggested by the purely physico-chemical character of the retinal processes. An analysis of the situation shows that in order to explain the behavior of color blind observers by the response functions thus obtained it is necessary to assume that "inner" dichromatism—such as deuteranopia—consists not in the absence of one of the fundamental responses, but in the failure of the mental organ to distinguish qualitatively between two of the fundamental responses. A satisfactory quantitative account is given of all the known forms of dichromatism.

A colorimetric system based on the "natural" or Young primaries compares favorably with the conventional I.C.I. system.

THE most fundamental fact about color perception is expressed by the Grassman law: If metameric lights are added to metameric lights, the sums are also metameric. The geometrical interpretation of this law is that the color sensations form an affine vector space. This space is three dimensional for the normal trichromatic observer and two dimensional for the dichromat.

The simplest way to account for Grassman's law is to assume the existence of a tri-receptor mechanism. Such an interpretation, most clearly expressed in the Young-Helmholtz theory, is not completely equivalent to the geometrical description. In the affine space any system of coordinates is equally good, whereas in the Young-Helmholtz theory three vectors, representing the fundamental sensations, have an exceptional position. These fundamental sensations, or primaries, are associated with the (fictitious) response of one of the color receptors with the exclusion of the other two. It is clear, therefore, that a reasonable determination of these "natural" primaries is of decisive importance for the theory.

Ever since Young and Helmholtz formulated their principle, several attempts were undertaken to determine the primaries corresponding to this theory. The calculations were either based on some definite assumptions on the hue of the primaries (so as to satisfy certain psychological requirements), or made so as to conform with the observed chromaticity confusions of different types of color-blind observers. This latter method was used by A. König[1] and revised recently by D. B. Judd[2] in accordance with the I.C.I. data and chromaticity confusion measurements of Pitt[3] and Hecht and Schlaer.[4] The calculations were performed on the assumption that dichromatism consists in the absence of one of the color responses. The obtained response functions agree fairly well with the observed luminosity functions of deuteranopes and protanopes. A serious difficulty arises from the circumstance that no luminosity is associated with the green and violet primaries of the trichromatic observer, in order to account for the fact that the luminosity function of a deuteranope is almost identical with that of a normal observer. This difficulty cannot be avoided unless additional assumptions are made, as, for instance, in the zone theory of G. E. Mueller. The chief objection which can be raised against such supplementary assumptions is that they are not accessible to quantitative treatment and deprive the Young theory of the utmost simplicity, which is no doubt its greatest attraction.

In this work, I consider the problem from a different angle. The Young theory of three

[1] A. König and C. Dieterici, Zeits. Psychol. **4**, 241 (1893).
[2] D. B. Judd, J. Opt. Soc. Am. **35**, 199 (1945).
[3] F. H. G. Pitt, Medical Research Council, "Report of the Committee on the Physiology of Vision," **14**, special Report Series No. 200 (London, 1935).
[4] S. Hecht and S. Schlaer, J. Gen. Physiol. **20**, 57 (1936).

receptors is primarily a physical hypothesis, and the quantitative consequences of the theory affect rather the physical than the subjective, psychical aspects of color vision. Therefore, when determining the Young primaries, it seems logical to start from assumptions which are physical in their nature instead of making assumptions on the hue of the primaries or drawing conclusions from the behavior of pathological observers. In Section 2 we will, in fact, be able to determine the primaries from some simple physical and geometrical assumptions, taking as basis the I.C.I. data for the average trichromatic observer.

Of course, a set of primaries which has been derived in this way, would be of little value if it were unable to account for the behavior of color-blind observers. This question will be raised and discussed in Section 3. It will be shown that every known type of dichromatism can be accounted for in an entirely satisfactory manner, provided that a certain distinction is made in the explanation of the two essentially different types of dichromatism called inner and outer color blindness by Mueller.

The usual interpretation—complete absence of one of the fundamental responses—is only retained for outer color blindness (protanopia and tritanopia). For inner color blindness, it will be assumed that the color receptor organs of these observers are functioning normally, but the sensations associated with two of the receptors are not discriminated in their mental apparatus. This pair of receptors is red-green for deuteranopes and red-violet for tetartanopes. No special assumption is made regarding the nature and location of the defect, but this is quite irrelevant for a quantitative phenomenological description of the process. It seems to me that this interpretation of deuteranopia is the only way out to account for both the luminosity function and color confusions of deuteranopes.

I will call the color system based on the Young primaries the "natural" color system. Before determining the primaries of this system, it will be useful to introduce as a new colorimetric concept, the three chromatic components of the light, which will allow us to draw a sharp distinction between objective and subjective elements in colorimetry and photometry in general. The definition of the chromatic components is given in Section 1, independently of any colorimetric system, their determination, however, will only be possible after finding the natural primaries.

The natural system offers an important basis for color specification. This aspect of the theory and some of its implications will be discussed in Section 4.

1. THE CHROMATIC COMPONENTS

In all the further discussions, the Young hypothesis will be regarded as the physical basis of color vision. We postulate the existence of three kinds of color receptor substances in the cones of the retina (for the sake of simplicity, we will call them the red, green, and violet receptors), which undergo some photo-chemical change from the action of light. The extent of this change is a function of the physical composition of the light. In order to obtain a measure of this change, say for the red color substance, we take a homogeneous red light of definite wave-length λ and intensity one, and consider its action on the red receptor as the unit. If a red light of the same wave-length λ and intensity I acts on the same spot of the retina, then we can take I as the measure of action of the corresponding light on the red receptor.

We can use an arbitrary light of some other wave-length λ' (or even an arbitrary inhomogenous light) to define a measure for the action on the red receptor. Evidently if the unit of intensity of the light λ' is chosen so that it shall act to the same extent on the red color substance, as the unit light λ, we arrive at the same definition of measure independently of the light chosen. To elucidate this important point more elaborately: Suppose two physically different lights L_1 and L_2 act to the same extent on the red-receptor elements of a certain part of the retina. That means that they are, as physical causes of the sensation, equivalent from the point of view of the red receptor. They remain equivalent even if the eye is in a different state of adaptation, that is, if the number of reactive molecules of the "red" photo-chemical substance and the state of fatigue of the optical nerves varies. Suppose we change the intensity of L_1 and L_2 in the same proportion, let us say, to $2L_1$ and $2L_2$. Then they still remain equivalent as the physical

causes of the process, and their "red action" (and the resulting sensation pertaining to the red receptor) will be equal again.

In the same way as described for the case of the red receptor, we define a measure of action on the green and violet receptors and thus attribute three numbers to an arbitrary light. Since the units are chosen arbitrarily, we submit them to the additional requirement that achromatic light, as represented by the equal-energy spectrum, should have equal measure numbers. We call the three numbers thus defined the *chromatic components* (in particular the red, green, and violet components) of the light.

Of course, there is a certain amount of sensation corresponding to a definite magnitude of the chromatic components, manifesting itself in the perception of a certain brightness. We can attribute a certain brightness to each of the three components, and these partial brightnesses are functions of the respective components. We could take as a measure of the partial brightness or luminance an arbitrary function of the chromatic components; perhaps some logarithmic function would best account for the phenomena of contrast sensitivity. To conform, however, with the customary measures of brightness, we take the chromatic components themselves as measures of the partial brightnesses, multiplied by appropriate proportionality constants.

These constants cannot be chosen arbitrarily, since we are endowed with the peculiar ability to compare the brightness of lights of different colors, that is, to make a statement on the equality of brightnesses pertaining to two different color receptors.

Here arises a peculiar situation which is greatly responsible for the weakness of the visual photometric concepts. For, what is the meaning of the judgment, that a red and violet light are of equal brightness? Certainly it does not mean that the same number of elementary processes of the red and violet photo-chemical substances are taking place in area and time unit, since they are different chemical processes in some very indirect connection with the sensation evoked by them. Such a judgment, that the exciting of the violet and red receptors results in a sensation of equal magnitude, is a fairly subjective one, it is remarkable indeed that such a judgment is possible at all. Now, suppose that we have a light with red component F_r, and one with violet component F_v, so that the red and violet brightnesses corresponding to F_r and F_v should be equal for a certain observer. Is it true that if we take a light with red component $2F_r$ and one with violet component $2F_v$, the corresponding red and violet brightnesses will still remain equal for the same observer? We have no reason to assume the validity of this statement, and in fact it is not strictly true even in the region of photopic vision (see reference 5). But then if we have defined say the red luminance as proportional to the red component of the light, we

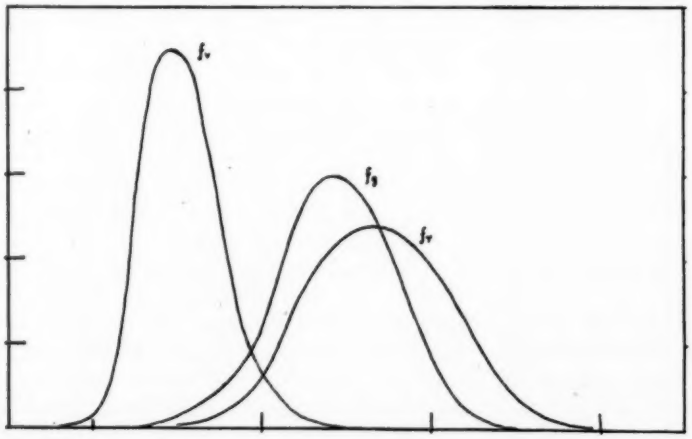


FIG. 1.

[5] Dresler, Das Licht **7**, 203 (1937).

have no right to define the green and violet luminances as strictly proportional to the respective components, but by some other function, which ought to be determined experimentally. In defining, however, the partial luminances as proportional to the respective components, and the resulting luminance as the sum of the partial luminances, although a more or less serious error is committed, we remain in conformity with the usage of photometrists who assume the additivity law of luminances. It must be kept in mind, however, that it is only the chromatic components which are strictly additive. This is the essence of Grassman's law.

Up to now we have considered the chromatic components as merely fictitious quantities, which, although they have a well established physical significance, are not measurable directly. I will show now that one can determine them as soon as the natural primaries are known in some way.

Let us perform color-mixture measurements in the usual way, that is an arbitrary light is matched against a mixture of three fixed independent primaries. The units of the primaries should be chosen so that equal-energy light should be balanced by a mixture of equal parts of the primaries, measured in these units. Suppose that the chromatic components of the three primaries (of unit quantity) are

$$f_r^{(1)}, f_g^{(1)}, f_v^{(1)}; f_r^{(2)}, f_g^{(2)}, f_v^{(2)}; f_r^{(3)}, f_g^{(3)}, f_v^{(3)}, \quad (1)$$

respectively. Then, if an arbitrary light is matched by $\bar{x}$, $\bar{y}$, $\bar{z}$ parts of the primaries, the chromatic components of the light are given by

$$\begin{aligned} F_r &= \bar{x}\cdot f_r^{(1)} + \bar{y}\cdot f_r^{(2)} + \bar{z}\cdot f_r^{(3)}, \\ F_g &= \bar{x}\cdot f_g^{(1)} + \bar{y}\cdot f_g^{(2)} + \bar{z}\cdot f_g^{(3)}, \\ F_v &= \bar{x}\cdot f_v^{(1)} + \bar{y}\cdot f_v^{(2)} + \bar{z}\cdot f_v^{(3)}. \end{aligned} \quad (2)$$

Thus the law of finding the chromatic components of a given light is the same as the transition to another trichromatic system with fictitious primaries, in which system the coefficients of our real primaries are given by (1). These fictitious primaries correspond, therefore, to fictitious stimuli which would excite only one of the color receptors, hence, they are identical with the natural primaries. Thus, if we had found the natural primaries, the trichromatic coefficients in this system would be identical with the chromatic components.

TABLE I.

λ	X	Y	ΔX	ΔY	$\Delta X/\Delta Y$
0.455	0.1827	0.0275			
0.450	0.1897	0.0214	−0.0126	0.0108	−0.86
0.445	0.1953	0.0167	−0.0096	0.0082	−0.86
0.440	0.1993	0.0132	−0.0071	0.0063	−0.89
0.435	0.2024	0.0104	−0.0056	0.0048	−0.86
0.430	0.2049	0.0084	−0.0043	0.0034	−0.79
0.425	0.2067	0.0070	−0.0032	0.0022	−0.69
0.420	0.2081	0.0062	−0.0024	0.0012	−0.50
0.415	0.2091	0.0058	−0.0017	0.0004	−0.24
0.410	0.2098	0.0058	−0.0013	0.0000	0.00
0.405	0.2104	0.0058	−0.0011	0.0000	
0.400	0.2109	0.0058	−0.0009	−0.0001	
0.395	0.2113	0.0059	−0.0007	−0.00015	
0.390	0.2116	0.0059½	−0.0006	−0.00015	
0.385	0.2119	0.0060½	−0.0005	−0.00015	
0.380	0.2121	0.0061			

One further remark: The trichromatic coefficients in a certain system are found solely through measurements of color matching, hence safe from the usual defects of heterochromatic photometry. In reality, such measurements do not require the concept of luminosity at all, since only one kind of judgment is made, namely, the complete identity of two sensations. Such a judgment can be made with considerable accuracy and has an objective meaning for a certain observer, at least in the case that only foveal vision is used, and that is the only one we are concerned with here. Indeed, measurements on color mixture are essentially physical measurements on the behavior of the photo-chemical substances of the observer.

As already mentioned, the chromatic components are independent of the adaptation of the eye; if a color match is obtained by an observer under favorable conditions of vision, then he will find it a match under any conditions of adaptation. On the contrary, the partial luminosities depend on the state of adaptation to a considerable extent; they are functions of the retinal adaptation of the three receptors. Essentially, this view is also expressed in the coefficient law of v. Kries.[6]

2. THE NATURAL PRIMARIES

Suppose we have found the natural primaries and calculated the chromatic components of

[6] J. v. Kries, "Die Gesichtsempfindungen," *Nagel's Handbuch der Physiology des Menschen* (Braunschweig, 1905), Vol. 3.

spectrally homogeneous, radiant energy of unit irradiance as a function of wave-length. In this way, three curves are obtained which may be regarded (apart from absorption in the ocular media) as the spectral-response curves of the color substances. In order to take into account the purely physical character of the chromatic components and response functions, we regard the color receptors as a physical system, most conveniently as simple resonators with respect to electromagnetic vibrations. The simplest form of a resonance function is a bell-shaped curve, converging exponentially to the wave-length axis in both directions. Hence, we will assume that each of the three response functions has this simple geometrical character.

The determination of the natural primaries will be essentially based on this one assumption; it almost suffices to determine the primaries uniquely. Among the possibilities which present themselves, I have chosen a system, the primaries of which are given in the I.C.I. system by the following trichromatic coefficients:

$$\begin{aligned} X_r &= 1.8517, & Y_r &= 0.6642, & Z_r &= 0.0000, \\ X_g &= -1.0630, & Y_g &= 0.3321, & Z_g &= 0.0020, \\ X_v &= 0.2113, & Y_v &= 0.0037, & Z_v &= 0.9980. \end{aligned} \tag{3}$$

Figure 1 shows the distribution functions obtained with these primaries. We have three bell-shaped curves, almost symmetrical about their maxima which lie at 0.567, 0.543, 0.447, respectively.

In order to make clear the amount of arbitrariness which is connected with the determination of (3), I will show the way by which one can arrive at these values. In the following much use is made of the I.C.I. table of trichromatic coefficients. A table can be found, for example, in the textbook of Moon,[7] p. 477.

First, we suppose tentatively that at the short-wave end of the visible spectrum only the violet receptor, and at the long-wave end only the red receptor, is excited to any substantial degree. This seems to be the only natural way to account for the fact that no hue changes with wave-length are being observed at both ends of the visible spectrum.

The I.C.I. coefficient table shows that at the long-wave end the coefficients of the spectral colors do not change appreciably and tend closely to a limit. We put for this limit (somewhat arbitrarily) $x=0.7360$, $y=0.2640$, $z=0.0000$. Denoting by X_r, Y_r, Z_r; X_g, Y_g, Z_g; X_v, Y_v, Z_v the trichromatic coefficients of the natural red, green, and violet primaries, we have

$$Z_r=0, \quad X_r/Y_r=0.7360/0.2640=2.7879. \tag{4}$$

We also see that Z_g must be small, since at $\lambda=0.570$, where the stimulation of the green receptor must be considerable, z is still very small.

Let us turn now our attention to the violet end of the spectrum. At first sight it would seem that the former argument applies to the violet primary too. A closer examination will show that this is not the case.

All the spectral blue colors under 0.450 act chiefly on the violet receptor, to a smaller degree on the green, and still more slightly on the red receptor.[8] As we proceed towards the violet extreme of the spectrum, we expect to find (relatively to violet) gradually decreasing stimulation of the red and green receptors. For example, in order to match blue 0.450, we have to add to blue 0.440 a small amount of a certain green stimulus which would act mainly on the green receptor and to a smaller degree on the red one.

Table I is extracted from the I.C.I. coefficient table; the I.C.I. coefficients of short-wave spectral lights have been reduced there to $z=1$. We find from the table that the stimulus which should be added to blue 0.440 in order to get blue 0.450, has the coefficients $\Delta X=0.1897-0.1993=-0.0096$, $\Delta Y=0.0214-0.0132=0.0082$, $\Delta Z=0.0000$ (here we made use of the fact that $Z_r=0$, and Z_g is comparatively small, at present altogether negligible). These are the coefficients of a mixture of the green primary with a relatively small amount of the red primary. The same calculation has been performed for all subsequent intervals of $\Delta\lambda=0.010$ and tabulated in Table I. We see that the function $\Delta Y/\Delta X$ has a fairly regular character down to $\lambda=0.420$, but changes

[7] P. Moon, *The Scientific Basis of Illumination Engineering* (1936).

[8] Here we have made use of the assumption that the red-response curve has a simple bell-shaped form and not the camel-back type as supposed, by some authors, on psychical grounds. (See for example, C. Ladd-Franklin, *Color and Color Theories* (New York, 1929).

its behavior under this wave-length. The most striking change is (the effect is small, nevertheless noticeable) that under 0.400 both coefficients of the difference stimulus are negative, that means the green-red stimulus must be subtracted and not added to the shorter wave-length light. This seems surprising at first, but can be easily understood if we remember that the extremities of all the three response curves must decrease exponentially. The above effect simply means that the violet curve approaches the wave-length axis more steeply than the other two curves—a behavior suggested anyhow by the general, steeper character of the violet curve. Thus, the spectral lights under 0.400 rather tend away from the fundamental violet than approach it.

Since the information obtained from the study of "difference stimuli" is scarce and not very accurate, we have to apply a trial and error method to find the exact coefficients of the violet primary. We make an approximate guess as to the coefficients of the violet primary somewhere around the spectral stimuli 0.380–0.400, then assign a trial value to Y_v/X_v which must be around -0.5 according to Table I, and determine the approximate course of the response functions. From this we infer the exponential character of the red and green functions at their extremities, and extrapolate them exponentially to the violet end of the spectrum. This will give the approximate amount of red and green components that must be subtracted from the extreme violet spectral stimuli in order to get the violet primary.

A careful study by this method has given the following values as the most probable I.C.I. coefficients of the violet primary:

$$X_v/Z_v=0.2117, \quad Y_v/Z_v=0.0037. \qquad (5)$$

If we subtract the violet primary (5) from spectrum lights around 0.430, we find for the difference stimulus a coefficient ratio $\Delta X/\Delta Y = -1.45$. On the other hand, the method of extrapolating the red and green curves shows that the ratio of red and green components in the vicinity of 0.420 is about 1:5. Subtracting this 20 percent red component from the above difference stimulus, we find the coefficient ratio X_g/Y_g to be approximately -3.1. The color-mixture data do not furnish sufficient information from which an exact value could be determined. The situation is not so bad, though, for if we try to assign to X_g/Y_g any value between -2.9 and -3.3, the resulting sets of response curves turn out to be almost identical, and the practical consequences are negligible. So, when deciding for a certain value of X_g/Y_g, we cannot do better than take a practical hint. If we try to determine a primary system, fixing X_g/Y_g somewhere between the above limits, we find that Y_r/Y_g is very near to 2. Now, it is preferable to have a simple value for Y_r/Y_g rather than for X_g/Y_g, since the former ratio expresses the relation of luminosities associated with the red and green primaries. We therefore put

$$Y_r/Y_g=2.000. \qquad (6)$$

Finally, we have to get a value for Z_g, which although small, is nevertheless not zero. Suppose we have to take α_r, α_g parts of the red and green primaries so as to obtain the sensation of green 0.570,* then the following equations hold (the trichromatic coefficients for 0.570 are taken from the I.C.I. table):

$$\alpha_r X_r+\alpha_g X_g=0.4441, \quad \alpha_r Y_r+\alpha_g Y_g=0.5547, \quad \alpha_g Z_g=0.0012.$$

Eliminating X_r, Y_r by (4) and (6), we have from the above equations

$$\alpha_g(2.788\,Y_g-X_g)=1.10,$$

and

$$Z_g/2.788\,Y_g-X_g=0.0012/1.10=0.0011. \qquad (7)$$

Through (4)–(7) the trichromatic coefficients of our new primaries are determined, save for a proportionality factor. This may be fixed by the additional requirement that equal-energy light shall have equal components, that is,

$$X_r+X_g+X_v=1, \quad Y_r+Y_g+Y_v=1, \quad Z_r+Z_g+Z_v=1.$$

---

* It is supposed here that the violet action is negligible at 0.570. There are two reasons for this assumption. First, the steep exponential character of the violet curve at the long-wave end is hereby preserved, and secondly the course of the z function of the I.C.I. coefficients from 0.570 upwards can be excellently accounted for if we attribute the z values solely to the green receptor. In the computation of (3), we have taken 0.0010 for the right-hand side of (7) in order to account for a small amount of violet at 0.570, found by extrapolation.

TABLE II.

λ	ξ_r	ξ_g	ξ_v	$1/\zeta_r$	$1/\zeta_v$
0.380	0.0020	0.0031	1.0000	1.55	0.0031
0.390	0.0018	0.0032		1.78	0.0032
0.400	0.0015	0.0034		2.27	0.0034
0.405	0.0014	0.0036		2.60	0.0036
0.410	0.0012	0.0039		3.18	0.0039
0.415	0.00105	0.0043		4.08	0.0043
0.420	0.0011	0.0052		4.81	0.0052
0.425	0.0014	0.0072		5.10	0.0072
0.430	0.0020	0.0099		4.86	0.0099
0.435	0.0031	0.0140		4.57	0.0140
0.440	0.0045	0.0195		4.30	0.0195
0.445	0.0063	0.0265		4.20	0.0265
0.450	0.0087	0.0358		4.12	0.036
0.455	0.0118	0.0477		4.05	0.048
0.460	0.0165	0.0639		3.87	0.064
0.465	0.0240	0.0863		3.59	0.086
0.470	0.0387	0.1238		3.20	0.124
0.475	0.0650	0.1837		2.83	0.184
0.480	0.1108	0.2814		2.54	0.281
0.485	0.1886	0.4385		2.32	0.438
0.490	0.321	0.693		2.16	0.693
0.494	0.480	1.000	1.000	2.08	1.000
0.495	0.491		0.9045	2.04	1.105
0.500	0.515		0.568	1.94	1.76
0.501	0.519		0.519		
0.505	0.5355		0.358	1.87	2.80
0.510	0.5555		0.219	1.80	4.56
0.515	0.578		0.130	1.73	7.70
0.520	0.600		0.079	1.67	12.7
0.525	0.624		0.052	1.60	19.2
0.530	0.649		0.0355	1.54	28.2
0.535	0.6755		0.0236	1.48	42.4
0.540	0.704		0.0152	1.42	66.0
0.545	0.736		0.0093	1.36	107
0.550	0.773		0.0055	1.29	180
0.555	0.815		0.0031	1.22	317
0.560	0.867		0.0016	1.15	630
0.565	0.929		0.0008	1.08	
0.570	1.000	0.9915	0.0004	0.992	
0.575		0.906		0.906	
0.580		0.815		0.815	
0.585		0.722		0.722	
0.590		0.628		0.628	
0.595		0.537		0.537	
0.600		0.452		0.452	
0.605		0.374		0.374	
0.610		0.306		0.306	
0.615		0.248		0.248	
0.620		0.200		0.200	
0.625		0.1615		0.161	
0.630		0.129		0.129	
0.635		0.102		0.102	
0.640		0.0795		0.0795	
0.645		0.0612		0.0612	
0.650		0.0472		0.0462	
0.655		0.0365		0.0365	
0.660		0.0285		0.0285	
0.665		0.0233		0.0233	
0.670		0.0191		0.0191	
0.680		0.0124		0.0124	
0.690		0.0077		0.0077	
0.700		0.0062		0.0062	

The solution of this system of equations is given by (3).

The trichromatic coefficients of an arbitrary color in the natural system, having coefficients X, Y, Z, in the I.C.I. system, are expressed by the following formulas:

$$\begin{aligned} F_r &= \ 0.2513X+0.8049Y-0.0562Z, \\ F_g &= -0.5027X+1.4015Y+0.1012Z, \\ F_v &= \ 0.0010X-0.0028Y+1.0018Z. \end{aligned} \tag{8}$$

These are the chromatic components of the light.

We could obtain unified coefficients by dividing F_r, F_g, F_v with $F_r+F_g+F_v$, but it seems more advantageous to refer the coefficients to the greatest coefficient as unit, that is, to divide them by $\max(F_r, F_g, F_v)$. The usefulness of this mode of expression will be seen in Section 4. We will call the coefficients, ξ_r, ξ_g, ξ_v, thus obtained, the normalized red, green, and violet components of the light or stimulus. Table II contains the normalized components of spectral colors, computed from the I.C.I. table by means of (8). For later purposes, the functions $1/\zeta_r=F_g/F_r$, $1/\zeta_v=F_g/F_v$ are also tabulated there.

The chromatic components of homogeneous spectrum lights of equal energy are obtained by applying transformations (8) to the standard distribution coefficients in the I.C.I. system (see Moon, reference 7, Appendix G, p. 595). They are tabulated in Table III and represented graphically in Fig. 1.

In the I.C.I. system, the primaries were chosen so that no luminosity was associated with the x and z primaries. In our case, positive luminosity is associated with all the three primaries. The relative luminosities are given by Y_r, Y_g, Y_v in (3). We call them the luminosity coefficients of the fundamental sensations:

$$l_r=0.6642, \quad l_g=0.3321, \quad l_v=0.0037. \tag{9}$$

They are numerically equal to the partial luminous emittances in an equal-energy source of unit luminous emittance. Hence, $\frac{2}{3}$ of the luminosity of equal-energy sensation comes from the red receptor, and the contribution of the violet receptor is less than 0.4 percent;

$$l_v/l_g=0.0111, \quad l_r/l_g=2. \tag{9*}$$

If the chromatic components of an arbitrary color sensation are F_r, F_g, F_v, then the partial luminances are

$$L_r=l_r\cdot F_r, \quad L_g=l_g\cdot F_g, \quad L_v=l_v\cdot F_v, \tag{10}$$

and the total luminance is

$$L=L_r+L_g+L_v=l_r\cdot F_r+l_g\cdot F_g+l_v\cdot F_v. \quad (11)$$

If the components here are those from Table III, we obtain the standard luminosity values.

The chromatic components of an arbitrary light may be obtained from the spectroradiometric curve by means of equations

$$F_r=\int f_r(\lambda)\cdot W_\lambda\cdot d\lambda,$$

$$F_g=\int f_g(\lambda)\cdot W_\lambda\cdot d\lambda,$$

$$F_v=\int f_v(\lambda)\cdot W_\lambda\cdot d\lambda, \quad (12)$$

where $f_r(\lambda)$, $f_g(\lambda)$, $f_v(\lambda)$ are values taken from Table III, and W_λ is the spectral radiance in watts per unit solid angle, unit area of the source, and unit wave-length interval. The chromatic components are expressed here in units similar to candles per unit area, and the luminances, both partial and total, in candles per unit area are given by (10) and (11).

TABLE III.

λ	f_r	f_g	f_v	λ	f_r	f_g
0.380	0.000012	0.000018	0.006	0.575	0.9485	0.860
0.385	0.000021	0.000035	0.011	0.580	0.9305	0.759
0.390	0.000036	0.000064	0.020	0.585	0.903	0.652
0.395	0.000060	0.00012	0.036	0.590	0.867	0.545
0.400	0.00010	0.00023	0.068	0.595	0.825	0.443
0.405	0.00016	0.00042	0.115	0.600	0.775	0.3505
0.410	0.00025	0.00081	0.208	0.605	0.719	0.269
0.415	0.00041	0.0016	0.372	0.610	0.657	0.201
0.420	0.00071	0.0034	0.647	0.615	0.591	0.1466
0.425	0.00145	0.0075	1.035	0.620	0.5215	0.1045
0.430	0.0028	0.0137	1.39	0.625	0.447	0.0722
0.435	0.0050	0.0228	1.625	0.630	0.375	0.0484
0.440	0.0079	0.0341	1.75	0.635	0.311	0.0318
0.445	0.0113	0.0474	1.79	0.640	0.2535	0.0201
0.450	0.0154	0.0638	1.78	0.645	0.202	0.0124
0.455	0.0206	0.0835	1.75	0.650	0.1573	0.0074
0.460	0.0276	0.1068	1.67	0.655	0.1207	0.0044
0.465	0.0368	0.1320	1.53	0.660	0.0905	0.0026
0.470	0.0500	0.160	1.29	0.665	0.0662	0.00155
0.475	0.0678	0.192	1.044	0.670	0.0477	0.00091
0.480	0.0902	0.229	0.814	0.675	0.0346	0.00054
0.485	0.1163	0.2705	0.617	0.680	0.0254	0.00032
0.490	0.1494	0.3225	0.4655	0.685	0.0179	0.00017
0.495	0.192	0.391	0.353	0.690	0.0123	0.000095
0.500	0.246	0.478	0.2715	0.695	0.0086	0.00006
0.505	0.3165	0.591	0.2115	0.700	0.0062	0.00004
0.510	0.3985	0.716	0.157	0.705	0.0044	0.00002
0.515	0.4905	0.849	0.110	0.710	0.0032	
0.520	0.583	0.971	0.077	0.715	0.0022	
0.525	0.663	1.063	0.055	0.720	0.0016	
0.530	0.733	1.129	0.040			
0.535	0.7915	1.171	0.0275			
0.540	0.840	1.193	0.018			
0.545	0.879	1.194	0.011			
0.550	0.909	1.177	0.0065			
0.555	0.9335	1.145	0.0036			
0.560	0.950	1.096	0.0017			
0.565	0.958	1.031	0.0008			
0.570	0.9575	0.951	0.0004			

$$\Sigma f_r+\Sigma f_g+\Sigma f_v=21.37$$

3. COLOR BLINDNESS

Originally, Thomas Young inferred his principle of the three color receptors from observations made on color-blind individuals. On the other hand, our previous discussion shows that it is possible to determine the Young primaries in a reasonable way without any reference to color blindness. In fact, we have obtained system (3) by using only such data which pertain to the average trichromatic observer. In this set-up of the theory, it is indispensable to examine how well the theory fits in with experimental data on color blindness. A closer examination shows that whereas the behavior of protanopes is excellently reproduced by the theory, there are considerable discrepancies in the case of deuteranopes if we adhere to Young's original interpretation of color blindness. To avoid this difficulty, we have the choice either to alter the system of natural primaries, or to modify the interpretation of color blindness. I will follow the second course, since any essential alteration of system (3) would necessarily destroy its logical consistency.

A good agreement with experiment can be established if we make the simple assumption that the defect of deuteranopes consists in an inability to discriminate the hues associated with

TABLE IV.

	Protanope		Deuteranope		
	Calculated 1	From Pitt 2	Calculated 3	From Pitt 4	Normal 5
0.430	0.024	0.04	0.009	0.06	0.012
0.440	0.044	0.06	0.019	0.07	0.023
0.450	0.070	0.09	0.032	0.09	0.038
0.460	0.104	0.11	0.052	0.10	0.060
0.470	0.145	0.15	0.081	0.12	0.091
0.480	0.199	0.22	0.127	0.16	0.139
0.490	0.273	0.30	0.195	0.23	0.208
0.500	0.401	0.44	0.306	0.34	0.323
0.510	0.599	0.67	0.482	0.52	0.503
0.520	0.810	0.87	0.686	0.71	0.710
0.530	0.941	0.97	0.840	0.85	0.862
0.540	0.994	1.00	0.938	0.92	0.954
0.550	0.981	0.97	0.988	0.97	0.995
0.560	0.913	0.88	1.000	1.00	0.995
0.570	0.792	0.75	0.971	0.99	0.952
0.580	0.632	0.60	0.902	0.95	0.870
0.590	0.454	0.44	0.801	0.87	0.757
0.600	0.292	0.29	0.683	0.76	0.631
0.610	0.167	0.17	0.555	0.62	0.503
0.620	0.087	0.10	0.428	0.48	0.381
0.630	0.040	0.05	0.301	0.33	0.265
0.640	0.017	0.02	0.200	0.22	0.175
0.650	0.006	0.01	0.123	0.12	0.107
0.660	0.002	0.002	0.070	0.08	0.061
0.670			0.037	0.03	0.032

the red and green receptors. We can make this assumption without inquiring into the physiological nature of the defect, just as the Young theory could be established without knowing the exact nature of the photo-chemical processes associated with the three color receptors.

Classification of Color Blindness

A defective functioning of the color perceiving apparatus can be attributed to one or more of the following circumstances.

a. Deviations in the position and shape of the response curves. This is very common and usually not very well marked. In most cases it can be attributed to the pigmentation of the ocular media. This is but to be expected, since the photo-chemical substances in the color receptors are likely to be the same for most observers, hence their spectral-response curves ought to be the same. It should be noted, however, that anomalous trichromatism is usually connected with this defect to such a degree that it cannot be attributed to the effect of pigmentation.

b. Outer defect: The luminosity coefficients have abnormal values. A variation in the luminosity coefficients l_r, l_g, l_v is very common and cannot be considered as a defect within certain limits. Even one and the same observer is subjected to great fluctuations depending on the state of adaptation of his eye. In the extreme case of outer dichromats one of the luminosity coefficients is zero.

c. Inner defect: The difference in the appearance of primary sensations is not so well marked as for a normal observer. In other words, the defective observer cannot discriminate as many hues between two primary sensations as the normal observer. In the extreme case of inner dichromats two of the primary sensations cannot be discriminated at all.

Monochromatism—except total cone blindness—can be attributed either to the inner or to the outer defect, or to both.

Let us now discuss the single cases of dichromatism.

1. Protanopia or red blindness. It is caused by the non-functioning of the red receptor, that is $l_r=0$. The luminosity function of the protanope can be obtained from Table III and Eq. (11) by putting $l_r=0$, $l_v/l_g=0.011$, and dividing $f_g+0.011\cdot f_v$ by the maximum of its value, 1.200. Column 1 of Table IV is constructed in this manner. It is in good agreement with experimental data of Pitt,[3] see column 2. It is also in good agreement with measurements of McKeon and Wright,[9] and Nelson[10] on protanomalous trichromats, which suggests that these observers have very low red-luminosity coefficients (provided that their green response function is normal). The maximum of the luminosity function is at 0.543.

To find the neutral point of the spectrum, we have to observe that the normalized chromatic components of color temperature 5000°K are $\xi_r=1$, $\xi_g=1$, $\xi_v=0.87$ (see Table V in the next section). Thus, the 5000°K neutral point in the spectrum of the protanope is given by $f_v=0.87\cdot f_g$ which is satisfied at $0.495\frac{1}{2}$. Finally, the chromaticity confusion lines must be copunctal at the red primary. From (3), its unified I.C.I. coefficients are $x=0.736$, $y=0.264$, which is in sufficient agreement with the result of Pitt.

2. Tritanopia, or violet blindness. It is caused by the absence of violet function, that is $l_v=0$. The luminosity function of the tritanope must be practically the same as the normal curve, with a slight deficiency on the violet side, provided that the ratio l_r/l_g is normal ($=2$). There is some indication that on the average, the ratio is actually larger, since the observed luminosity function seems to be shifted to the long-wave side (maximum of the luminosity function at 0.560 instead of 0.555). The 5000°K neutral point is at 0.570 ($f_g=f_r$). The chromaticity confusion lines must be copunctal to the violet primary, that is, to the point with unified I.C.I. coefficients $x=0.174$, $y=0.003$. No experimental data are available to check this result.

3. Deuteranopia, or red-green confusion. The discussion of this case is not quite as simple as it was in the foregoing cases. To determine the chromatic components for the deuteranope, we have to know the relative contribution of the red and green receptors to the equal-energy sensation. The most obvious assumption would be to take them in the proportion of their normal trichromatic luminosity coefficients. This would

[9] W. M. McKeon and W. D. Wright, Proc. Phys. Soc. London **52**, 464 (1940).
[10] J. H. Nelson, Proc. Phys. Soc. London **50**, 661 (1938).

be too much of a restriction though, and we prefer to leave $\alpha = l_r/l_g$ undetermined at first. Denoting by f_y the response function of the—now indistinguishable—red-green primary sensation (the index y is an allusion to yellow), we have

$$f_y = (\alpha/1+\alpha)\cdot f_r + (1/1+\alpha)\cdot f_g, \qquad (13)$$

where f_r, f_g are taken from Table III. It should be noted here that in deriving (13) we have tacitly assumed that the additivity law is strictly valid when applied to the *different* receptors r and g. If the analysis of Section 1 is correct and deuteranopia is really an inner color blindness (hence, the receptors r and g *de facto* different), then this assumption is not correct. This would imply that Grassman's law is not strictly valid for the deuteranope. This question could be decided experimentally and perhaps is worth studying. Of course, no great deviations can be expected from the strict additivity law.

Assuming that (13) is strictly correct, we can calculate the lines of constant chromaticity from

$$F_v/F_y = \text{constant},$$

or, by (13),

$$F_v/F_g + \alpha\cdot F_r = \text{constant}.$$

Substituting (8) into this equation, we see that the lines of constant chromaticity are straight lines, copunctal at the point

$$F_v = 0, \quad F_g + \alpha\cdot F_r = 0,$$

or

$$0.0010x - 0.0028y + 1.0018z = 0,$$
$$(0.2513\alpha - 0.5027)x + (0.8049\alpha + 1.4015)y + (0.1012 - 0.0562\alpha)z = 0.$$

Writing here $z = 1-x-y$ and solving for x and y, we obtain

$$x = 1.001 - 1.004y, \quad y = 1 - 0.5\alpha/3.8 + 1.1\alpha.$$

Pitt found experimentally that the lines of constant chromaticity are almost copunctal at $x = 1.08$, $y = -0.08$. The fact that they are not strictly copunctal can be taken, perhaps, as an indication that the Grassman law does not strictly hold.

If we put $\alpha = 2.00$, which is the average trichromatic value of the ratio l_r/l_g, then we obtain $x = 1.00$, $y = 0.00$ as the intersection point of the lines. To make it equal to the Pitt value, we have to put $\alpha = 3.16$. This higher value of α is further supported by the fact that the luminosity function of the average deuteranope is slightly shifted to the long-wave side, as compared with the normal trichromatic observer. In Table IV, column 3, the luminosity function is calculated by means of the formula

$$(3.16f_r + f_g + 0.011f_v)/4.10.$$

The agreement with Pitt's data is well within experimental limits; see column 4.

The 5000°K neutral point is obtained from

$$\frac{3.16}{4.16}f_r + \frac{1}{4.16}f_g = \frac{1}{0.87}\cdot f_v,$$

which is satisfied at $\lambda = 0.500\frac{1}{2}$.

4. Tetartanopia, or violet-red confusion. This is a rare case of inner color blindness, which can be explained by the identity of the red and violet primary sensations. No chromaticity confusion data are available, so the relative strength of the sensations contributed by the red and violet receptors in equal-energy light cannot be computed as in the previous case.

If we put $\beta = l_v/l_r$ and denote by f_p the response function of the common red-violet sensation (p = purple), we have

$$f_p = (\beta/1+\beta)f_v + (1/1+\beta)f_r,$$

and the 5000°K neutral point of the spectrum is given by

$$f_p = \frac{1+0.87\beta}{1+\beta}\cdot f_g,$$

or

$$\frac{\beta}{1+0.87\beta}\cdot f_v + \frac{1}{1+0.87\beta}\cdot f_r = f_g.$$

This equation is generally satisfied at two different spots of the spectrum. It was found in the few cases hitherto studied[11] that the neutral points of tetartanopes lie around 0.580 and 0.470. These values are best approximated if we take $\beta = 0.135$, which furnishes 0.576 and 0.474 as neutral points. The disagreement is not so bad if we take into account the strongly impaired

[11] L. L. Sloan and S. M. Newhall, Am. J. Ophtal. **25**, 945 (1942).

ability of tetartanopes to distinguish hues at all, which obviously makes it difficult for them to locate the neutral points correctly. The value $\beta=0.135$ is 25 times the average trichromatic $l_g/l_r=0.0055$.

The lines of constant chromaticity must be nearly copunctal at the point satisfying

$$F_g=0, \quad F_r+\beta\cdot F_v=0,$$

or, putting $\beta=0.135$,

$$\begin{aligned} -0.503x+1.401y+0.101z&=0,\\ 0.251x+0.805y+0.079z&=0,\\ x+y+z&=1, \end{aligned}$$

which gives $x=-0.04$, $y=-0.10$. No data are available to check this result.

The luminosity function depends on the value of l_g/l_r, regarding which no *a priori* assumption can be made. Since the luminosity function of tetartanopes is reported to have a maximum around 0.560, it is safe to assume that l_g/l_r is certainly not greater than the normal value. This implies that both l_r and l_g have abnormally low values relative to violet, and the color vision of such observers must be strongly impaired. This would explain their resemblance to monochromats.

Theoretically, two more types of dichromatism are possible, namely green blindness and violet-green confusion. It seems that these cases have never been observed. It must be noted, however, that violet-green confusion, if existed, would be very similar to violet blindness. In fact, the available data on tritanopia suggest that it is rather a case of violet-green confusion than one of violet blindness.

4. SPECIFICATION OF COLORS

From the geometrical point of view, any trichromatic system is equally well suited for color specification. On the other hand, the Young-Helmholtz theory predicts an exceptional position for a system based on the three fundamental primaries. The following considerations should serve to demonstrate the exceptional role played by the natural system.

Suppose the trichromatic observer A performs color-mixture experiments in the usual way, by matching two half-fields of an illuminated screen. In another series of experiments, he performs color matching so that he places a slightly colored filter between his eyes and the observation screen. Let the system "observer A plus filter" be called the observer B. Let us compare the colorimetric systems set up by A and B.

Let the chromatic components of a spectral light λ (for the observer A) be F_r, F_g, F_v, then we obtain those for the observer B (measured in the same units as A) by multiplying F_r, F_g, F_v with the spectral-transmission factor $c(\lambda)$ of the filter. If we regard B as an independent observer, then we have to change the units of the chromatic components so that equal-energy light should have equal components for B, measured in the new units. Thus, the chromatic components of the light for the observer B will be, after adjusting the units,

$$F_r'=c(\lambda)\cdot\epsilon_r\cdot F_r, \quad F_g'=c(\lambda)\cdot\epsilon_g\cdot F_g, \quad F_v'=c(\lambda)\cdot\epsilon_v\cdot F_v,$$

where the factors ϵ_r, ϵ_g, ϵ_v are independent of λ.

It is advantageous to refer the chromatic components to the green component as unity. Writing

$$\zeta_r=F_r/F_g, \quad \zeta_v=F_v/F_g, \quad e_r=\epsilon_r/\epsilon_g, \quad e_v=\epsilon_v/\epsilon_g,$$

we have

$$\zeta_r'=F_r'/F_g'=e_r\cdot\zeta_r, \quad \zeta_v'=F_v'/F_g'=e_v\cdot\zeta_v.$$

Hence, the g-normalized components ζ_r', ζ_v' for the observer B are obtained from those of A by multiplying them with the λ-independent factors e_r, e_v. The reciprocals of the functions ζ_r and ζ_v for the average trichromatic observer are found in Table II.

The absorption of light in the ocular media has the same effect as a colored filter. Supposing that the color substances are the same for all the observers—with a possible exception of anomalous trichromats—we conclude from the above, that an individual color system, set up by a trichromat with the aid of an arbitrary set of primaries, can always be transformed so that the g-normalized trichromatic coefficients in the new (natural) system should be those from Table II, multiplied by certain unit factors e_r, e_v. This shows that the specification of spectral colors by means of the chromatic components has, in a certain sense, an invariant character, and is com-

TABLE V.

λ	ξ(g/v)	λ	ξ(v/g)	ξ(r/g)
0.380	0.0011	0.494	1.000	
0.390	0.0014	0.495	0.811	
0.400	0.0019	0.500	0.111	
0.405	0.0022	0.501	0.000	0.000
0.410	0.0027	0.505		0.276
0.415	0.0032	0.510		0.431
0.420	0.0042	0.515		0.515
0.425	0.0058	0.520		0.566
0.430	0.0079	0.525		0.603
0.435	0.0110	0.530		0.636
0.440	0.0150	0.535		0.667
0.445	0.0203	0.540		0.699
0.450	0.0274	0.545		0.733
0.455	0.0364	0.550		0.772
0.460	0.0481	0.555		0.815
0.465	0.0638	0.560		0.867
0.470	0.0885	0.565		0.929
0.475	0.127			
0.480	0.192			
0.485	0.308			
0.490	0.548			

From 0.570 to 0.700, the values for ξ(g/r) are identical with those for ξ_g, $1/\zeta_r$ in Table II.

TABLE VI.

Temp. °K of Panckian radiator	ξ_r	ξ_g	ξ_v
290	1.000	0.009	0.000
599	1.000	0.124	0.000
997	1.000	0.354	0.006
1497	1.000	0.561	0.045
1896	1.000	0.673	0.110
2355	1.000	0.764	0.211
2842	1.000	0.835	0.336
3493	1.000	0.901	0.505
4790	1.000	0.984	0.819
6486	0.877	0.916	1.000
9979	0.645	0.711	1.000
23950	0.486	0.563	1.000
∞	0.416	0.495	1.000

pletely identical, apart from the two coefficients e_r, e_v, for each individual observer. The unit factors e_r, e_v depend only on the pigmentation of the ocular media, and can be taken into account for each individual observer. Thus, the situation in this respect is similar to a color system which uses homogeneous spectrum lights as primaries, like the WDW system of Wright.[12]

a. Specification of Colored Light

We are used to describing colors (the term is used here to denote colored light) in the psychological terms hue, saturation, and brightness. All these concepts have "objective" counterparts which can be formulated particularly clearly if we employ the natural system.

Of course, it must be borne in mind that an objective specification can only have a limited bearing on the subjective appearance of colors. This latter depends largely on adaptation, surround illumination, and various psychical factors which cannot be included in an objective specification system. Therefore, the concepts "dominant wave-length" and "purity" (or color content), which are the objective equivalents of hue and saturation, have an essentially schematical character.

If we could produce a stimulus which would excite only one of the color receptors, we could perceive the purest conceivable sensation of red, green, or violet. But apart from these three definite fundamental hues, all others could only be evoked by exciting two of the receptors simultaneously. Hence, if we had a stimulus which would excite, in a certain proportion, the violet and green receptors only, but not the red, we had the purest conceivable sensation of blue. Blue light however, even spectral blue, always excites all the three receptors simultaneously. Take, for example, spectral blue λ=0.480. From Table II we see that its normalized components are: $\xi_r=0.110$, $\xi_g=0.281$, $\xi_v=1.000$. We can split up this light mathematically into two parts, namely, a stimulus with components $F_r=0.110$, $F_g=0.110$, $F_v=0.110$, and one with components $F_r'=0.010$, $F_g'=0.171$, $F_v'=0.890$. The first one is an achromatic stimulus, while the second one is a fictitious stimulus which would excite only the green and violet receptors in the given proportion. If we could produce this stimulus, it would evoke the purest possible sensation of blue 0.480. (Of course, only approximately, since the hues do not remain unchanged if we add or subtract equal-energy sensation.) Therefore, we may say that the sensation of spectral blue 0.480 has an achromatic content of 11.0 percent and a color content of 89.0 percent, the latter being a pure blue with normalized components $\xi_r=0$, $\xi_g=0.171/0.890=0.192$, $\xi_v=1$. For brevity we

[12] W. D. Wright, Trans. Opt. Soc. London **30**, 141 (1928–29).

introduce the notation $\xi(g/v)=0.192$, which number is characteristic of the color $\lambda=0.480$.

Generally, if we have the normalized components of a color, and ξ_1 denotes the smaller, ξ_2 the bigger component, then the achromatic content is given by $100\cdot\xi_1$, the color content or purity by $100(1-\xi_1)$, and the characteristic number is obtained from the formula $\xi_2-\xi_1/1-\xi_1$. The characteristic numbers for spectral colors are tabulated in Table V. They serve to find the so-called dominant wave-length of an arbitrary color, without the use of a color triangle, provided the chromatic components are known.

As seen from Table III, the greater part of the spectral colors is not pure, but contains a more or less considerable amount of achromatic sensation. This is a deviation from the ordinary view that spectral colors are of 100 percent purity, but is more in accordance with the implications of the Young-Helmholtz theory. A similar formulation of color purity was originally proposed by Exner,[13] but the idea cannot be developed consistently if not based on the natural primaries.

Table VI shows the normalized components of the light from Planckian radiators at various temperatures, computed from data of Judd.[14]

b. Object Colors

A similar specification can be applied to object colors. Let the surface of the object be illuminated by a standard equal-energy illuminant, and the chromatic components of the reflected light be F_r, F_g, F_v. Let the (equal) chromatic components of the light reflected by a white surface under the same illuminant be F. Then the color of the surface can be characterized by the normalized components

$$\eta_r=F_r/F, \quad \eta_g=F_g/F, \quad \eta_v=F_v/F.$$

If η_1 is the smallest, η_2 the middle, and η_3 the greatest among the above three components, then $100\cdot\eta_1$ shall be defined as the "white percentage," $100(\eta_3-\eta_1)$ as the "color percentage" and $100(1-\eta_3)$ as the "black percentage" of the color. The sum of the white, black, and color percentage is 100 percent. The characteristic number for the dominant wave-length is given by $\eta_2-\eta_1/\eta_3-\eta_1$.

[13] Exner, Sitz. Akad. Wiss. Wien **119**, 233 (1910).
[14] D. B. Judd, J. Opt. Soc. Am. **23**, 366 (1933).

5. CONCLUSION

The Young-Helmholtz theory, if adequately formulated, can give a wholly satisfactory account of the objective phase of color perception. More than that cannot be required from the theory. Being essentially a physical hypothesis, its direct implications must remain strictly within the physical domain. It is the more surprising that several well-known objections to the theory are based on arguments which affect the subjective phase of color perception.

One of the most common objections is the following: The original form of the Young hypothesis—and also the one given in this paper—requires that the three fundamental sensations of a normal observer should be violet, green, and red. This seems to be in contradiction with the fact that the two hues observed by a protanope and deuteranope in the spectrum are blue and yellow instead of violet and green or red—a fact also corroborated by the testimony of observers with monocular protanopia.[15]

It must be emphasized in this connection that there is no forcible reason, why we should regard the fundamental sensations pertaining to different observers as actually identical. If the explanation of deuteranopia given in Section 3 is correct (and the experimental evidence seems to support it) then this problem is put in a rather acute form. The deuteranope observes the same hue, regardless whether his r or g receptor is excited. What, then, is this sensation "in reality," red, green, or yellow? There seems to be no sensible answer to this question, unless we view it from a practical angle. The testimony of the deuteranopes themselves cannot settle the issue. Objectively, they perceive in the spectrum hue 1 and hue 2, and calling them blue and yellow, they just express a resemblance to the hue of common object colors whose names they have learned from trichromatic observers. Obviously, their assertion has no bearing on the question of actual appearance of hues.

From a practical point of view, it is advantageous to assume that achromatic color has the same appearance for every observer. (Postretinal adaptation will accomplish this to a considerable

[15] W. Dieter, Zeits. Sinnesphysiol. **58**, 73 (1927).

extent for observers with monocular protanopia.) Under this assumption, we have to identify the two fundamental hues of protanopes and deuteranopes with two complementary hues of the trichromatic observer. The least confusion is caused by identifying them with a certain blue and orange-yellow of the spectrum. Which wavelength we actually choose to represent these sensations is merely a matter of convention that cannot decide for or against the Young-Helmholtz theory.

A similar criticism can be applied to attempts of identifying the Young primaries with psychically unitary or invariable hues. Even if we admit that the fundamental sensations are invariable hues, it cannot be predicted what happens to the hue of a homogeneous spectrum light if the intensity is changed. We must remember that spectral lights always excite at least two—but generally all the three—receptors. By changing the intensity of such a stimulus, the hue of the sensation is affected by various psychical factors which are not directly connected with the Young primaries.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 38, NUMBER 4 APRIL, 1948

Visibility on Cathode-Ray Tube Screens: The Effects of Light Adaptation*

R. M. HANES AND S. B. WILLIAMS
Psychological Laboratory, The Johns Hopkins University, Baltimore, Maryland

(Received October 20, 1947)

An intensity modulated cathode-ray screen of the PPI type used in radar depends on differential brightness discrimination for detection of weak signals. The brightness quantities have been largely unknown, because of photometric difficulties, but it has commonly been assumed that dark-adaptation is required for efficient visual detection. Therefore, an experimenal test was undertaken of the effects of light- and dark-adaptation on the contrast threshold and on time required for detection of suprathreshold signals appearing on a typical screen (P7 phosphor). Adapting brightnesses ranged from zero to 2000 millilamberts (ml.). By special photometry, the four screen intensities used were estimated as ranging from 0.0001 ml. to 0.22 ml. Length of exposure to the brightest adapting light and the degree of contrast between signal and background were also varied. The lowest contrast thresholds and the shortest detection times were found to occur when the test and adapting illuminations were equal or nearly so. When the two illuminations were unequal, the loss in sensitivity was roughly in proportion to their difference. The optimal field brightness of the phosphor screen used was on the order of 0.1 ml. With this field the level of adaptation of the eye can be a relatively insignificant factor in radar detection, even of weak signals. Increasing values of signal contrast allow progressive reduction in detection delay after exposure to the highest brightness.

INTRODUCTION

IN a previous report[1] the general problem of visibility of signals on cathode-ray screens was discussed. In that paper it was pointed out that for certain screens—and probably for all—there exists an optimal brightness level, which is determined by two factors: the differential sensitivity of the eye and the differential sensitivity of the cathode-ray tube. The present study is concerned with the first factor, the eye's differential sensitivity.

The relatively low brightness levels at which radar screens are operated, compared to the high brightness to which the eye may be exposed, makes imperative some knowledge of the nature and magnitude of any possible effects of such exposure. We need to know, for example, whether there is some optimal brightness level to which the eye should be adapted before observation is begun. On the other hand, what

* This research was carried out under Contract N5ori-166, Task Order I, between Special Devices Center, Office of Naval Research, and The Johns Hopkins University. This is Report No. 166-I-32 under that contract. The authors are indebted to Mr. Ellwood King for engineering assistance throughout the study and to Dr. N. R. Bartlett and Dr. C. T. Morgan for suggestions in writing the paper.

[1] S. B. Williams, N. R. Bartlett, and E. King, "Visibility on cathode-ray tube screens: screen brightness" (in preparation).

is the order of magnitude of the effects of exposure to bright lights on the sensitivity of the eye? We need to know how much the threshold is changed by varying amounts of brightness. How long does it take an eye which has been adapted to a high brightness to see a very low contrast? How much does this contrast have to be increased to allow immediate detection? These questions in radar work are raised in connection with the detection of signals or pips of different brightnesses. Despite the prodigious number of studies on dark adaptation in the literature, relatively few have been concerned with the problems raised here.

Several experiments by Craik[2,3] are the most relevant to the present problem. Using eight test field brightnesses, he found that the lowest threshold (about one percent) occurred when the test and adapting illuminations were equal. Sensitivity was considerably less when measured at an illumination differing greatly from that to which the eye had been adapted. For adapting intensities of four e.f.c. and above, the eye shifted its region of maximal sensitivity to correspond to the adapting intensity in such a way that deviations from it produced proportional losses in sensitivity. Thus Craik could fit all of his data roughly to one family of curves, the members of which differed only in their maxima and not in their shape. This means that no matter at what absolute brightness level the eye is adapted, the differential threshold would be about 100 percent if measured at 1/1000 of the adapting brightness, or 2 percent if measured at or near the adapting intensity, and so on. This is an important fact to keep in mind in designing lighting arrangements for radar rooms or for any room in which cathode-ray screens are to be viewed in such a way as to require the most effective use of the eye's differential sensitivity. Sensitivity requirements are paramount in intensity modulated screens such as the Plan Position Indicator but are probably of less importance with *A*-scopes.

The above principle holds only for illuminations of a few footcandles and above. An exact value for the lower limit cannot be set for all cases because it depends somewhat on the size of the object viewed. It will, in general, fall between 0.1 and 10 e.f.c. Below this limit Craik reports that contrast sensitivity is greater when the test field is slightly brighter than the adapting illumination. This is because the brighter test field takes advantage of the eye's increasing sensitivity at low illuminations, provided the difference is not great enough to introduce a subjective glare effect or to require any appreciable time for bright adaptation. This does not mean that it would be better viewing practice, in radar, for example, to have the eyes adapted to lower than test field illuminations. No advantage can be gained from any factor other than the prevailing brightness of test field. Furthermore, as the present report will clearly show, too much dark adaptation may impair sensitivity.

Data obtained by Wright[4] show very little change in contrast sensitivity with high preexposure (adaptation) illuminations. The difference between these and Craik's results can be explained in terms of the range of test-field brightnesses employed. Wright's study was restricted to the middle intensities where contrast sensitivity is nearly constant anyway. In other words, his study was limited to a range of brightnesses where the Weber fraction is constant.

The course of rod dark adaptation as influenced by the intensity and duration of light adaptation was studied by Haig.[5] He found that the curve of dark adaptation following varying degrees of light exposure exhibits: (a) a decrease in slope and (b) a displacement to the right on the time axis. In general these changes can be produced either by raising intensity or prolonging exposure. From this and from adaptation time curves of the type published by Hecht, Haig, and Chase[6] we can infer that recovery to low brightness, near the level of scotopic vision, would require proportionately more time than recovery to moderate or high brightness levels. Within limits the data of Blanchard[7] would appear to confirm this. Haig determined recovery times for a wide range of adapting intensities but for only one selected test field and for one degree of

[2] K. J. W. Craik, J. Physiol. **92**, 406 (1938).
[3] K. J. W. Craik, Brit. J. Psychol. **29**, 252 (1939).
[4] W. D. Wright, J. Physiol. **83**, 466 (1935).
[5] Charles Haig, J. Gen. Physiol. **24**, 735 (1941).
[6] S. Hecht, C. Haig, and A. M. Chase, J. Gen. Physiol. **20**, 831 (1937).
[7] J. Blanchard, Phys. Rev. **11**, 81 (1918).

contrast. Apparently only Blanchard[7] has systematically varied degree of contrast, and his data are restricted to rather low adaptation levels, never higher than 0.1 millilambert (ml.). Nevertheless, in shifting from this value to lower brightnesses, recovery time was nearly inversely proportional to the degree of contrast. An experiment by Simonson, Blankstein, and Carey[8] employed a procedure similar to part of the work in this study. They measured the length of time required to reach a given threshold as a function of length of exposure to the adapting brightness. However, their longest exposure time was only about thirty seconds. An experimental comparison of recovery times for a wide range of test field intensities simultaneously with a wide range of adapting intensities does not appear to be in the scientific literature.

Even with these excellent studies at hand we are not able to solve adequately the problems confronting the radar operator as outlined above. In the first place, the nature of the discrimination to be made in the type of radar viewing screen used here is different from any called for in the laboratory experiments. This point will be made clear in the description of the apparatus which follows. In addition, while absolute brightness values are always very hard to determine, the problem with radar is unusually difficult due to the fact that the background brightness is constantly varying. (This problem will also be treated below.) Even though we were able to obtain exact quantitative data in brightness values from other studies, there would still remain the problem of translating these values into the voltage units used in radar. For the reasons mentioned it was deemed necessary to provide a systematic practical study on cathode-ray tube screens.

The actual experiment consisted of four parts: (I) the effect of the length of exposure to a constant high brightness on the time required to see a given contrast; (II) the effect of exposure to different intensities on the differential threshold; (III) the effect of the intensity of light adaptation on the time required to detect a given contrast; and (IV) the degree of contrast necessary to overcome the effects of light adaptation. Each of these parts will be described in detail. But before the treatment of the adapting methods and procedure of each part, a description of the apparatus will be given.

[8] E. Simonson, S. S. Blankstein, and E. J. Carey, Am. J. Ophthal. 29, 328 (1946).

APPARATUS

Cathode-Ray Tube

The test field was the circular face of a No. 7BP7 tube, schematically pictured in Fig. 1. The tube is the familiar Plan Position Indicator (PPI) with the radial scan. The screen is the P7 phosphor, which allows relatively long persistence of images. It fluoresces blue (peak emission at about 435 mμ) and phosphoresces yellow (peak emission at about 570 mμ). The signal appears as a pip of brighter light on the background, which in turn may be varied from bright yellow to a very dim neutral gray by adjusting the bias on the control grid. The tube was operated at voltages specified by the manufacturer; all circuit voltages were highly stabilized and closely monitored. No visible noise (background clutter) appeared on the screen; the only significant excitation of the phosphor came from the focused beam of the electron gun. Stray external light was excluded by a shield.

Measurement of Field Brightness

Direct measures of field brightness are inherently difficult with a screen of this kind, not only because the lower levels lie below the limits of most photometers and photoelectric cells but also because the field is heterogeneous and cyclically changing with scanning rate. Furthermore there is the problem of defining the psycho-

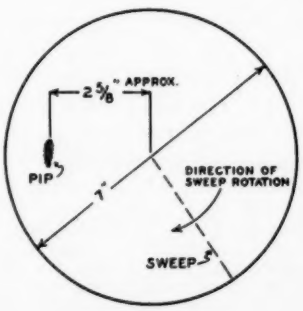


FIG. 1. Test field.

TABLE I. Brightness in millilamberts of each of four test fields and the corresponding CRT bias necessary to produce it. Probable brightness values are estimated for 0 and 2 volts from VRI by extrapolation of measurements made 1/6th second after excitation; the value for 4 and 6 volts from VRI are actual measurements.

Volts from VRI (22 v)	Brightness of field of pip
0	0.0001
2	0.0005
4	0.022
6	0.22

logically effective field of the pip. We here define it as the pip's immediately adjacent background; this background can be expected to represent various points in the scanning cycle depending on CRT bias.[9]

Calculated values for field brightness appear in Table I. Physical measurements were made with a special photomultiplier tube (931A) and d.c. amplifier circuit. The brightness values can be regarded as reasonably exact. Under the conditions used, a CRT bias of 22 volts is the Visual Reference Intensity commonly used at this laboratory as an anchor point to the brightness scale.[1] VRI is the bias at which the sweep line is barely visible to the dark-adapted eye, a very stable value for a given observer. Deviations from VRI were measured with a voltmeter; positive numbers mean *less* bias and therefore a brighter field.

Signal

The signal was a simulated radar echo made by a signal generator and sent, after video amplification, through an attenuator to the control grid of the cathode-ray tube. It produced a pip on the screen at approximately the halfway point from the center to the periphery, as shown in Fig. 1. Radial and angular position could easily be varied. To some extent pip size could be varied also. The signal used throughout this experiment was two microseconds long and twenty degrees in azimuth. This produced a pip which subtended a retinal angle, at twelve inches viewing distance, of three degrees in length by seven minutes in width.

[9] A. L. Sweet and N. R. Bartlett, "Visibility on cathode-ray tube screens: Signals on a P-7 screen seen at different intervals after excitation," J. Opt. Soc. Am. (in press). This paper will show that the optimal portion of the scanning cycle is on or near the fluorescent sweep line for dim screens (bias 17 to 20 volts) but is several degrees behind the sweep line for brighter screens.

TABLE II. Brightness in millilamberts of pip and background, and the ratio between them.

Test field brightness	Brightness of pip	Contrast ratio
0.0001	0.0061	60.00
0.0005	0.0022	3.40
0.022	0.0286	0.30
0.22	0.2600	0.18

The intensity of the pip was varied by attenuating a standard signal. The standard signal reference intensity was maintained from day to day by the following procedure. Using the VRI screen brightness, the signal voltage was adjusted until the standard observer reported a threshold pip on the screen when there was exactly thirty decibels of attenuation of the signal. This was approximately an eight-volt signal.

Adaptation Light

The surface used to adapt the eye to high brightnesses was a large ground glass screen illuminated from behind by two 300-watt Mazda lamps. This screen was placed adjacent to, and at right angles to, the cathode-ray screen. The color temperature of the lamps was about 2850 degrees Kelvin. Measured in its brightest region with a Macbeth illuminometer, the glass surface had a brightness of 2000 millilamberts. Lower adapting brightnesses could be obtained by Wratten neutral filters. A goggle was made to hold the filter and at the same time fit over the left eye of the observer. The size of the opening in the goggle was such that the visual field was filled by the adapting light.

PART I. THE EFFECT OF LENGTH OF EXPOSURE TO A CONSTANT HIGH BRIGHTNESS ON THE TIME REQUIRED TO SEE A GIVEN CONTRAST

This part of the study was actually carried out some time earlier than the rest of the work and the data used for improving the experimental techniques of later studies. However, some of the data on exposure time are regarded as worthy of report.

Our interest lay chiefly in determining two things: (a) the way in which delay in detection

varies with increasing length of light adaptation, and (b) the length of exposure needed to give a maximal delay. The literature dealing with the latter point is rather sketchy, some writers claiming that light adaptation to a bright source is over in one or two minutes and others believing that such adaptation requires considerably more time before a steady state is reached.

The independent variable was the length of time the observer's eyes were exposed to the 2000 millilambert adapting light. Immediately after exposure, the eyes were turned to the test field (the cathode-ray screen), and the time required to detect a signal pip ten decibels above threshold voltage was read with a stop-watch. This signal was considerably above threshold, perhaps by as much as three or four times the amount of the threshold uncertainty interval.

A parameter in the experiment was the test field or screen brightness. Recovery time was determined for each of the four brightnesses mentioned above (0.22 ml. to approximately 0.0001 ml.). The ten-decibel contrast ratio was not the same, in corresponding brightness units, for all four test fields. As nearly as can be calculated with present measuring instruments, these contrast ratios (taken as the brightness of the pip minus the brightness of the test field divided by the brightness of the latter) are as stated in Table II. (It should be emphasized that these brightness values, as well as the others used throughout this article, are for a given point in the pip-scan interval. Both absolute and relative brightnesses change in value beyond this point because of the nature of the phosphor decay curve.)

Because of this limitation the recovery times for the four test fields are not directly comparable for purposes of interpretation of the visual adaptation process. They are comparable for the interpretation of required signal powers in radar.

Adaptation Procedure

Eight different lengths of exposure to the 2000-millilambert source were used: $\frac{1}{4}$, $\frac{1}{2}$, 1, $1\frac{1}{2}$, 2, 3, 4, and 5 minutes. This adapting source was so located that it filled the observer's entire field of view. Binocular vision was used in this part of the experiment, as contrasted with the rest of the study, in which monocular viewing was employed.

Before initial light adaptation for each series of exposures the observer waited for two minutes in darkness. He was then exposed to the adapting source for $\frac{1}{4}$ minute, after which he looked immediately at the test field. After reporting the pip as detected, the observer waited in the dark for one minute before the next period of light adaptation. This interval plus the time required in detecting the pip was judged sufficient to allow the eye to reach approximately the same level of adaptation in all cases before light adaptation was started. This procedure (i.e., light adaptation for the selected length of time, detection of the signal, one-minute dark interval) was continued until the entire series of exposure times, from $\frac{1}{4}$ to 5 minutes, had been completed. Exposure times were also taken in reverse order, starting with 5 minutes adaptation and ending with $\frac{1}{4}$ minute, and in random order. Since no systematic differences due to order were obtained, the assumption that the level of adaptation was very nearly the same before each exposure seems to be substantiated.

Results

The results obtained from the above procedure are shown in Table III and depicted graphically in Fig. 2. The table gives the means of the recorded times for the combined results of two observers. Also presented are the ranges of the scores and the number of observations in each

TABLE III. Time for detection (in seconds) at different lengths of exposure to a constant intensity for four test field brightnesses.

Test field brightness (millilamberts)	Statistic	Length of exposure to 2000 ml. (minutes) $\frac{1}{4}$	$\frac{1}{2}$	1	$1\frac{1}{2}$	2	3	4	5
0.0001	Mean	15.3	19.2	27.8	35.4	46.2	48.5	43.1	43.7
	Range	14.4	15	15	15	27	27	27	21
		19.2	26.4	45	54	66.6	63	61.2	63
	N	8	8	8	8	8	8	4	4
0.0005	Mean	9.5	12.5	15.1	27.4	23.8	25.4	27.6	32.7
	Range	7.8	8.4	8.4	15	8.4	8.4	8.4	15
		15	15.6	24	37.8	49.8	47.4	55.8	63
	N	10	10	10	6	10	10	10	6
0.022	Mean	6.9	7.0	6.6	6.8	7.2	8.0	8.0	7.4
	Range	6	6.6	5.4	6	6	5.4	5.4	4.8
		7.8	7.8	7.8	7.8	7.8	11.4	11.4	9.6
	N	4	4	4	4	4	4	4	4
0.22	Mean	15.9	19.3	34.1	36.6	37.7	55.3	56.5	62.6
	Range	7.2	8.4	19.8	19.8	14.4	21.0	15.0	37.8
		25.2	33.6	63.0	66.0	67.8	87.0	90.0	85.8
	N	10	10	10	10	10	10	10	10

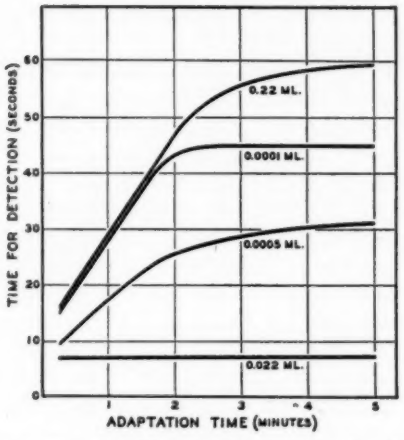


FIG. 2. Time for detection of pip as a function of length of exposure to 2000 millilamberts for four test field brightnesses. Curves have been drawn to give the best fit by eye to the plotted points. For empirical values see Table III.

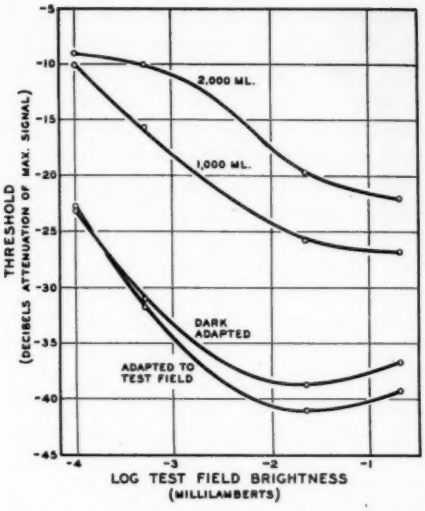


FIG. 3. Pip threshold as a function of test field brightness for four adapting brightnesses.

case. There were considerable individual differences at the longer exposure times except at the 0.022 test field brightness where detection times were minimal for all lengths of exposure. One of the observers had consistently shorter detection times than the other. In spite of these differences, however, the curves for each subject level off at the same point and the shape of the individual curves is very similar. All curves were fitted by eye to the empirical points.

As the graph shows, the curves for the darkest and brightest test fields very nearly coincide for part of their length. In other words, the longest detection times were obtained when these two brightness levels were used, but for different reasons. Detection in the darkest field suffers because the field is maximally different from the adapting brightness. The brightest field shows poor detection, in spite of its approximation to the adapting brightness, because of the unfavorable contrast ratio. The 0.0005 test field was somewhat better for detection than either. The 0.022 field gave minimal detection times no matter what length of light adaptation had been used. Its superiority over the brightest field is because of its more favorable contrast ratio; it is superior to the dimmer fields because it is nearer the adapting brightness used.

Inspection of the range of obtained values given in Table III will show that the differences in mean detection times are not significant after one or two minutes of light adaptation. This indicates that increased lengths of light adaptation beyond several minutes for a 2000-millilambert source have very little or no effect for the type of test used in this study. Adaptation, for all practical purposes, is complete in about two minutes.

It should be noted that a detection time of about seven seconds is a minimum. This fact is caused largely by the scan interval (six seconds) and the point at which light adaptation was terminated for each trial. In order to insure that the observer would have an opportunity to see the pip on its first presentation during any trial, adaptation was stopped each time about four or five seconds before the sweep passed over the pip location. This interval plus the reporting time accounts for the high minimal value.

The fact that none of the curves (except that for the 0.022 ml. test field) shows a perfectly horizontal course after the two-minute exposure time is not considered important. As Table III shows, the mean detection time is greater in some cases for shorter adaptation times than it is for the longer ones. These discrepancies are no doubt due to the limited number of observations taken for each point.

Practical implications are obvious. Even a very brief exposure to a high brightness (sky brightness, for example) may seriously impair a radar operator's ability for detection immediately afterward. The effect is minimized when looking at brighter scopes and stronger signals.

PART II. THE EFFECT OF ADAPTATION TO DIFFERENT INTENSITIES ON THE DIFFERENTIAL THRESHOLD

Differential threshold measurements were made after varying degrees of light adaptation in order to determine the maximal effects that such adaptation can have on very weak signals or pips. The four test field brightnesses given in the description of the apparatus were used. For this part of the experiment the independent variable was the intensity of the adapting source. Four adapting brightnesses were used for each test field. Three of these values (darkness, 1000, and 2000 millilamberts) were the same for all four test fields. The fourth adapting brightness, however, was the test field itself and therefore varied with the brightness of test field employed. This method was used to obtain absolutely minimal values for each test condition.

Adaptation

Exposure time to each intensity was sufficient to insure adaptation. From the data obtained in Part I, five minutes was deemed more than enough adaptation time for the higher adapting brightnesses. The adaptation time was increased proportionately for the lower intensities. For the 1000- and 2000-millilambert levels, then, the initial adaptation time was five minutes. Adaptation to this level was maintained by never allowing the eye to be removed for longer than two seconds, just time enough to make a threshold judgment of the pip on the test field. Light adaptation was renewed for two minutes between successive exposures to the test field, so that virtually constant adaptation was assured. When the cathode-ray screen itself was used as the adapting source, an initial adaptation time of ten minutes was allowed; renewal was not necessary, of course, inasmuch as the eye looked at the field continuously. For adaptation to darkness, the observer spent 25 minutes wearing red dark-adapter goggles in ordinary room illumination followed by five minutes in complete darkness. He then looked at the test field for two seconds, made a threshold judgment, and renewed his dark adaptation for ten minutes before the next two-second exposure.

Threshold Determination

In every case the threshold is an "appearance" threshold, that is, it was approached from below. On successive presentations the strength of the signal was increased from a value below threshold by one decibel steps. This is really one-half of the method of limits, i.e., its ascending series only. (Descending series are not feasible because of slow phosphor decay.) Therefore, each threshold is probably a fraction of a db higher than it otherwise would be. Three threshold determinations were made for each condition by the single observer.

A red fixation light was used for the two darker test fields. It was so placed that the image of the pip fell about 10 degrees temporally from the fovea.

Results

The data yield a family of curves which can be plotted in two ways. In Fig. 3 the mean threshold is shown as a function of the test field brightness,

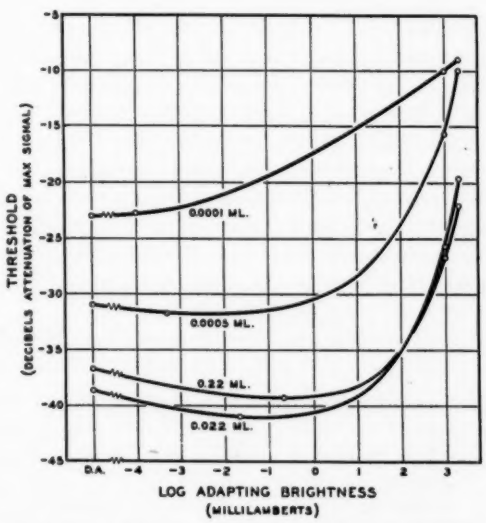


FIG. 4. Pip threshold as a function of adapting brightness for four test field brightnesses.

TABLE IV. Visibility threshold for four degrees of light adaptation at each of four screen brightnesses. Visibility thresholds recorded as decibels attenuation of a standard signal. The mean threshold for each condition is expressed both in db and, in parentheses, as a brightness contrast. Contrast ratios were computed from brightness measures and their extrapolations, as explained in the text.

Screen bright-ness	Dark-adapted eye		Adapted to screen		Adapted to 1000 ml.		Adapted to 2000 ml.	
	recorded values	mean	recorded values	mean	recorded values	mean	recorded values	mean
0.0001	23		23		11		10	
	24	23.0	23	22.67	10	10.0	9	9.0
	22	(3.10)	22	(3.20)	9	(385.0)	8	(860.0)
0.0005	31		32		16		11	
	31	31.0	32	31.67	16	15.67	10	10.0
	31	(0.72)	31	(0.66)	15	(12.0)	9	(60.0)
0.022	38		41		26		20	
	39	38.67	41	41.0	26	25.67	20	19.67
	39	(0.12)	41	(0.09)	25	(0.60)	19	(1.40)
0.22	37		39		27		21	
	36	36.67	39	39.33	26	26.67	22	22.0
	37	(0.08)	40	(0.06)	26	(0.24)	23	(0.45)

with adapting brightness as the parameter. In Fig. 4 the threshold is plotted as a function of adapting brightness, with test field brightness as the parameter. In both figures the ordinate is the number of decibels attenuation of the standard signal voltage. The complete list of empirical threshold values in Table IV reveals the small amount of variability among the data.

Adapting intensities of 1000 and 2000 millilamberts produced a marked rise in the threshold. Dark adaptation caused a slight rise, except for the darkest test field. Except in the case of the dark-adapted eye looking at the darkest test field, the threshold was lowest when the eye was adapted to the brightness of the test field itself. In other words, when a dark-adapted eye looks at a brighter screen, its threshold is higher than after it has become adapted to the screen. The loss is small (two db of signal power) but is significant. When we remember that the brighter screen produces the better visibility, we can see that complete dark adaptation could be detrimental to radar detection.

That the brighter screens are better for visibility is attested not only by the data of Part I but by the present results as well. This fact is particularly evident in Fig. 4. In this connection we may wonder why the very brightest screen is less favorable for visibility than the next brightest, as is the case. The answer lies in the cathode-ray tube, not in the eye. The screen brightnesses selected for this experiment more than span the range of linear response of the screen's emission. Consequently, in the brightest screen, the tube is near its limit and its increment of light output per unit of voltage change is reduced.

These results are more or less in accord with the previous studies mentioned in the introduction. They correspond directly with the optimal conditions reported by Craik. Inasmuch as he used eight test brightnesses, twice our number, his conclusions are the more surely grounded. Our results, however, are an interesting confirmation of the phenomenon with different test stimuli. The troublesome matter of equating voltage units (in which our thresholds are expressed) to illumination units prevents our concluding, as Craik could do very well, that the loss in sensitivity is proportional to the difference in the adapting illumination.

The practical conclusion, then, from all the data available, is that for optimal visual sensitivity in viewing cathode-ray screens which require differential brightness discrimination, the eyes should be adapted to illuminations approximating that of the viewing screen. Since the optimal viewing screen is moderately bright for P7 phosphors, the eyes should not be completely dark adapted.

TABLE V.

Test field brightness (ml.)	Pip brightness (ml.)	Contrast ratio
0.0001	0.0007	6.00
0.0005	0.0010	1.00
0.022	0.0248	0.13
0.22	0.2380	0.08

PART III. THE EFFECT OF THE INTENSITY OF LIGHT ADAPTATION ON THE TIME REQUIRED TO DETECT A GIVEN CONTRAST

It was the purpose of this part to the study to determine how much difference there is in the effects of various degrees of light adaptation. This time the independent variable was brightness of the adapting source, while in Part I it was the length of exposure to a constant brightness. The measure of effect was the same as for the first part of the experiment: the time required to detect a pip of a given contrast. However, because of the considerable differences in contrast

TABLE VI. Time for detection of pip for various degrees of light adaptation at each of four screen brightnesses. (Target contrast: 3 db above threshold.)

Screen brightness (millilamberts)	Dark-adapted eye		0.20 ml.		2.0 ml.		6 ml.		200 ml.		1000 ml.		2000 ml.	
	rec. val.	mean	rec. val.	mean	rec. val.	mean	rec. val.	mean	rec. val.	mean	rec. val.	mean	rec. val.	mean
	Time for detection (seconds)													
0.0001	5	5	11	10	(19)*	37	(41)	59	449	441	776	799.5	901	968
	5		9		36		60		433		823		1035	
0.0005	5	5	5	5	5	5	15	15.5	240	244.75	493	469.67	(483)	621.5
	5		5		5		16		236		447		663	
									231		(293)		(423)	
									272		469		580	
0.022	15	14	5	5	5	5	7	6	79	81	140	141.67	(400)	194.5
									90		(209)		(393)	
	13		5		5		5		74		141		199	
											144		190	
0.22	7	6	5	5	6	5.5	5	5	20	19	98	85.5	138	140.5
	5		5		5		5		18		73		143	

* Note: Figures in parentheses were not used for various reasons such as equipment variability, lack of certainty by observer, etc.

ratios encountered in Part I with a signal ten db above threshold, it was decided to minimize this effect by using a three-db contrast in this part of the study. For this reason and because monocular viewing was used here, no attempt will be made to compute intensity-duration equations. The differences in technique make difficult the equation of the two variables. Haig[5] has already shown that there is a reciprocal relation between duration and intensity in their effects on the rate of rod adaptation. Very likely a similar relation holds for mixed rod and cone adaptation or for cone adaptation alone, both of which are represented in the present experiment.

The same four brightnesses already described were used as the test fields. Using the threshold values determined in Part II, we increased the intensity of the pip by a constant voltage increment (three db), just enough to make it certain of detection. As nearly as can be determined the contrast ratios for the pips at the four test field levels are as shown in Table V.

Adaptation

Seven exposure intensities were used to adapt the eye: darkness and 0.2, 2, 6, 200, 1000, and 2000 millilamberts. Again the data from Part I were used to determine adequate adaptation times. For the dark-adapted state the observer wore red dark-adapter goggles for 25 minutes in ordinary room illumination followed by five minutes in complete darkness. He then turned immediately to the test field, and the length of time necessary before the pip could be detected was determined. After detection of the pip, renewal of dark adaptation was allowed for ten minutes. A second observation immediately followed the renewal of adaptation. In some cases, if there was very little or no variability in the two successive recorded times (as indicated in Table VI), only two observations were made for those particular values of test field and adapting brightness. If considerable variability was noted, as shown by the table, several more observations were made.

TABLE VII.

Adapting brightness level	Initial adaptation time (min.)	Time for renewal of adaptation (min.)
Dark adaptation	30	10
0.2 ml.	10	5
2 ml.	10	5
6 ml.	10	5
200 ml.	5	2
1000 ml.	5	2
2000 ml.	5	2

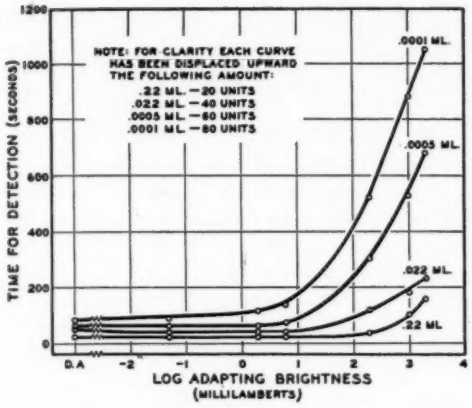


FIG. 5. Time for detection as a function of adapting brightness for four test field (scope) brightnesses.

After detection time for the dark adapted state had been determined, a short rest period was given; observations were then started to determine values for the next higher adapting level. The procedure was the same for each adapting brightness; only the adaptation times were varied. Table VII shows the time given to initial adaptation and to renewal for each brightness level.

All determinations for each test-field brightness were made on the same day, that is, a complete function was determined for each screen brightness in one experimental session. While a long session of this type is quite tiring to the observer, it was deemed necessary in order to minimize equipment variability for any one function. It is not believed that the length of the experimental session was so long as to affect the results seriously.

A red fixation light was used, as in Part II, for the two dimmer scopes. As mentioned earlier, it was so located that the image of the pip fell about 10 degrees temporally from the fovea.

Results

The obtained data, as given in Table VI, have been plotted in two ways. Figure 5 shows the time for detection as a function of the adapting brightness, with test field brightness as the parameter. In Fig. 6 direction time is plotted as a function of test field brightness, with adapting intensity as the parameter. The detection time in each case is given in seconds from the end of adaptation. This method of recording detection time allows a minimal time of not less than four or five seconds, since a few seconds are required for the observer to remove the goggles and turn to view the test field. In other words, an ordinate value of five seconds on the graphs represents immediate detection of the pip. The minimal time is less than that found in Part I of this study because a scan interval of only two seconds was used in this case.

It is obvious from the figures that increasing intensities of visual adaptation produce greater delays in detection. This tends to be true for each of the screen brightnesses but is considerably more marked when the test field is dim than when it is bright. Adapting illuminations up to one millilambert have little or no measurable effect on recovery times for the brighter screens. This quite surely is because the test field more nearly approximates the adapting intensity and not much recovery of sensitivity is required. Since the brightness of the pip's immediate background on the screen was never higher than about 0.22 millilamberts, it is evident that a considerable difference between test and adapting intensities can be tolerated without noticeable recovery delays.

Reference to Fig. 5 reveals that the recovery time curve for the 0.022-ml. screen departs significantly from its minimal value at about 10 ml.; that for the 0.22-ml. screen at about

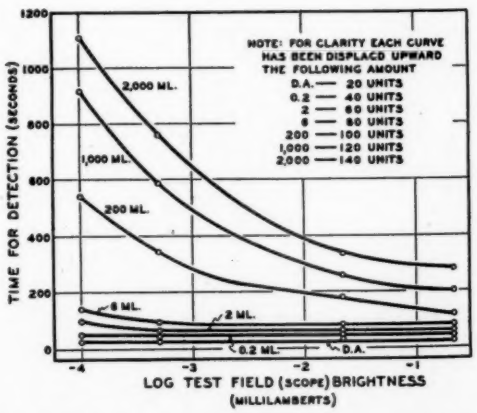


FIG. 6. Time for detection of pip as a function of test field brightness for seven adapting brightnesses.

100 ml. or one log unit higher. For both screens the departure from minimal delay occurs when the adapting brightness is almost three log units greater than that of the test field. This conclusion cannot be generalized to include the dimmer screens, and it looks as if it might not hold for them. The curve for the dimmest screen apparently departs from a minimal value in considerably less than the three-log unit difference mentioned above.

In the study by Haig[5] mentioned in the introduction there are some data comparable to this part of the present experiment. His method was to measure the time required to see a stimulus patch 0.75 log units above the absolute (dark adapted) threshold after exposure to intensities up to 4700 millilamberts. His data follow an *S*-shaped function up to this brightness. However, the first part of the function is very similar to the uppermost curve in Fig. 5, that for the dimmest screen. In fact, his curve crosses the two-log unit vertical at about five minutes, the same as does the curve here.

It is to be noted that dark adaptation does not produce any measurable recovery time for our dimmest screen. There is an appreciable though small effect on the threshold, as was shown in Part II, but it apparently is not large enough to persist for five seconds. However, this effect is slightly evident in the case of the two brighter screens.

As mentioned above, it is not possible to make a general comparison between the intensity data and the duration data because of the differences in technique used in these two parts of the experiment. It is worth while to note, nevertheless, that at the points where the degree of light adaptation was the same in these two cases (five minutes adaptation to 2000 millilamberts), the data for Part I show much shorter delays than were obtained here. While most of this discrepancy can easily be explained in terms of the brighter pips used in Part I, part of the difference is probably due to the use of binocular vision in that part of the study. A short check experiment seemed to substantiate this conjecture; binocular delay time was shorter than monocular under the same experimental conditions.

Another point is the difference in the optimal screen brightness between this part and the first part of the experiment. Here, as the graphs show, the brightest screen gave the shortest delays; in Part I the next brightest screen was found to be best. The answer again (see discussion under Part II) lies in the cathode-ray tube and the phosphor. Because of the limits on the tube's emission, a ten-db increase in voltage does not give for the brightest screen in Part I the same proportional increase in brightness units that a three-db increase does in Part III. In other words, when the brightness has passed the extent of linear response of the screen's emission, the increment of light output per unit of voltage change is reduced.

The foregoing data have some meaningful implications for viewing practice in radar. In looking at a radar display for information, the operator will be able to see weak signals almost immediately on a moderately bright screen no matter whether his eyes have been previously light adapted or not, provided the light adaptation was no higher than ten millilamberts. A similar degree of light adaptation would seriously delay his perception of weak signals on a dim scope. Therefore, the allowable limits of illumination for radar rooms ought to be rather low for dim scope operation but can be extended considerably when brighter screens are used.

PART IV. THE DEGREE OF CONTRAST NECESSARY TO OVERCOME THE EFFECTS CAUSED BY LIGHT ADAPTATION

Thus far, in the first three parts of the experiment, we have presented data on the effects of various degrees of light adaptation in the detection of signals on the cathode-ray screen. The fourth part of the study was designed to determine how much increase in pip brightness is required to compensate for the detrimental effects of previous light adaptation on detection time.

After the eye has been adapted to a given illumination level, if it is required to see something placed against a background of a different brightness it will have to adjust or adapt itself to the new level. This adaptation process is not instantaneous and the greater the brightness difference between the two levels the longer will be the time required. On the other hand, com-

TABLE VIII. Brightness equivalents of pips produced by various signal voltages. Signal intensity given in decibels above threshold, which is about 41 db of attenuation of an 8 volt signal. ΔI is the difference between pip and background in millilamberts (calculated) and $\Delta I/I$ is the contrast ratio.

Signal (db)	ΔI	$I+\Delta I$	$\Delta I/I$
3	0.0028	0.0248	0.13
6	0.0040	0.0260	0.18
10	0.0066	0.0286	0.30
15	0.0128	0.0348	0.58
20	0.0264	0.0484	1.20
25	0.0550	0.0770	2.50

plete adaptation to the test field is not necessary merely for an object in the field to be seen. The eye can detect the object long before complete adaptation is attained provided the object is very much brighter than its background. Presumably there should be a proportional relationship involved such that detection time should shorten as the degree of contrast is increased. This hypothesis was tested in this part of the study.

Only one test field brightness, the 0.022-millilambert level, was employed here. This is a representative condition and is the optimal brightness level for the CR tube employed. The eye was adapted to each of the seven illumination levels given in Part III; the adaptation time for each level was the same as indicated there. Six different degrees of contrast were used. The first or lowest degree of contrast was the same as that used in Part III, that is, 3 db above the visual threshold contrast determined in Part II. The other five contrasts, in terms of decibel units, were 6, 10, 15, 20, and 25 decibels above the visual threshold. The calculated values for these contrasts in brightness units are shown in Table VIII.

Adaptation procedure was the same as in Part III: the eye was adapted to each brightness level, detection time was determined, the eye was adapted to the next level, etc. Inasmuch as there was very little or no variability in the recorded times, only two trials were run at each brightness level for any given degree of contrast. When the variability between successive trials warranted it, an additional reading was taken (see Table IX). No observations were made at the contrast values and adapting intensities left as blank spaces in the table. It should be noted that the detection times listed at the left end of each row in the table, i.e., at the various contrasts, are minimal values (see Part II). It was assumed that all detection times to the left of these values, that is, at the lower adapting brightnesses, would also be minimal. These

TABLE IX. Time for detection of pip for various degrees of contrast between target and background. (Constant scope brightness: 0.022 ml.)

Contrast ratio	Dark-adapted eye		0.20 ml.		2.0 ml.		6.0 ml.		20.0 ml.		100.0 ml.		200.0 ml.	
	rec. val.	mean	rec. val.	mean	rec. val.	mean	rec. val.	mean	rec. val.	mean	rec. val.	mean	rec. val.	mean
	Time for detection (seconds)													
0.13	15	14	5	5	5	5	7	6	79	81	140	141.67	199	194.5
	13		5		5		5		74		144		190	
0.18	5	5	5	5	5	5	5	5	13	12.5	90	89	135	131.5
	5		5		5		5		12		88		128	
0.30							5	5	5	5	33	49.33	101	92
							5		5		61		83	
0.58									4	4.5	5	5	58	54
									5		5		50	
1.20											5	5	20	19.5
											5		19	
2.50											5	5	5	5

assumed values are filled in on the graphs plotting the functions.

It was unnecessary to use a fixation light with the test field used in this part of the study.

Results

The complete set of data obtained is given in Table IX with the means of the recorded times at each degree of contrast for each adapting brightness. The detection time in each case is given in seconds from the end of adaptation. This, as pointed out in Part III, allows a minimum time of not less than four or five seconds. In other words, a recorded time of about five seconds represents immediate detection of the pip.

Figure 7 shows the mean detection time plotted as a function of the adapting brightness for each degree of contrast. In Fig. 8 these degrees of contrast are given on the abscissa. This plot shows a little more clearly the very large differences between weak and strong signals so far as detection is concerned.

For several reasons much of the work in this series has been concerned with weak signals or dim pips. In the first place, in order to keep contrast relatively constant for different test field brightnesses, it was necessary to use targets near threshold, since brightness values were not known at the time the experiment was designed. Secondly, we were interested in the extreme limits of harmful effects produced by high adapting intensities, and these would, of course, be most evident with weak signals. This graph portrays very well just how the results of the rest of the study must be modified to suit the particular problem at hand. If very strong signals or pips are coming through, the range of tolerable adapting intensities can be much greater than that given in the preceding parts of the experiment. Nevertheless, for the sake of precaution, it is safest to follow those recommendations which will give optimal viewing conditions. The eyes should not be exposed to very high brightnesses immediately before scope work is started.

From the graphs it is evident that for the lowest degree of contrast both light and dark adaptation cause an increase in detection time. (This is the same condition reported in Part III.)

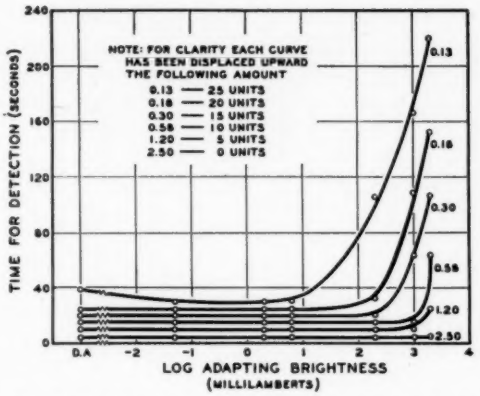


FIG. 7. Time for detection of pip as a function of adapting brightness for six degrees of contrast. (Test field brightness: 0.022 ml.)

Dark adaptation, however, causes a delay of only fourteen seconds whereas adaptation to high brightnesses may cause delays up to more than three minutes. The delay becomes less as the relative intensity of the pip increases. It is to be noted that even the weakest signal can be seen almost immediately even after adaptation to as much as six millilamberts. Slightly stronger signals can be detected readily after adaptation to as high a value as 200 millilamberts. *Even well-lighted rooms rarely have surfaces as bright as 6 ml., unless the walls are of high reflectance; 2000 ml. corresponds to the brightness of blue sky away from the sun.*

The replot of the data in Fig. 8 is chiefly of practical value as a guide to radar operation. It shows at a glance the supra-threshold signal strength required to overcome the effect of light exposure. With the highest exposure, 2000 ml., a signal of 25 db above threshold is required if it is to be detected immediately. Exposures requiring little or no recovery time are those closest to the scope intensity; namely, 0.2, 2, and 6 ml. Exact values can be read from Table IX.

OTHER EXPERIMENTS

Although the data of the present study are based largely on the observations of one eye, there is no reason to believe that they present a picture of the relative effects of light and dark adaptation which would be untrue for binocular

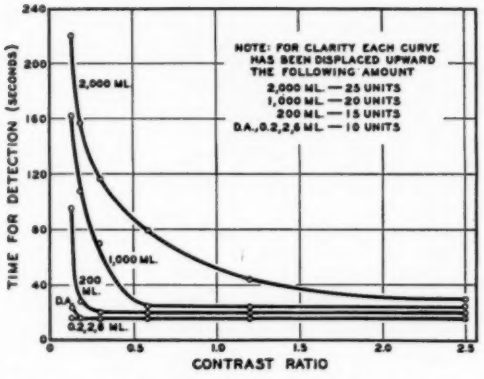


FIG. 8. Time for detection as a function of degree of contrast for seven adapting brightnesses. (Test field brightness: 0.022 ml.)

vision or for viewing under less controlled and less standardized conditions, inasmuch as adaptation processes probably are retinal functions. Furthermore, we have reason to believe that similar results would be obtained by other eyes; a few of the data were repeated with another observer with essentially similar results. In addition, a preliminary experiment in which four relatively untrained observers participated showed that adaptation intensities over a narrow range had little effect on visibility but that what slight effect did exist was in line with the results of the major experiment.

A brief summary of this preliminary experiment follows. Four college students took turns determining appearance thresholds of pips of unknown location on the same kind of cathode-ray screen (7BP7). Two screen brightnesses were used: Dim, very close to the VRI, and Bright, very close to the brightest screen, in Table I. Two adaptation levels were used: adaptation to diffuse room illumination which produced eight e.f.c. on the gray panel adjacent to the *CR* screen, and thirty-minute adaptation to the same illumination with red dark-adapter filters covering the eyes. Viewing was binocular; viewing distance was not fixed but was approximately ten to fifteen inches. Pips were detected by their location and reported exactly as in radar practice. Viewing was begun immediately following adaptation to either high or low intensity and pips were reported within six seconds, usually within two to four seconds. (It took six seconds for a complete circular scan, and since the pip's location was changed from trial to trial, the time of actual pip sighting was distributed over the six second interval.) In short, the experimental conditions were very similar to those which might exist in operational radar, except that both a higher and a lower adaptation level were employed than would likely exist in most radar rooms.

The results show that eyes looking at the bright *CR* screen were slightly handicapped by having been dark adapted, the difference between the eight e.f.c. and the red goggle pre-exposures being 2.5 decibels for threshold signals. This difference is on the borderline of statistical significance. There was even less difference when looking at the dim *CR* screen. The dark adapted eyes had an advantage of one decibel over the light adapted eyes, a difference which is statistically insignificant. The conclusion we draw is that within the ranges of ordinary room illumination, the adaptation level of the eyes is an insignificant factor. Dark adaptation may, with rather bright screens, be slightly harmful to target detection. Light adaptation, up to a few foot candles, certainly is not damaging to detection on bright screens but may be slightly damaging if the screen is very dim.

Since monocular vision had been used in the four main experiments, there might be some question as to a possible displacement of the right ends of the curves (in Parts II, III, and IV) downward if binocular viewing were to be employed. Such a supposition seems probable in the light of data obtained from another brief preliminary experiment in which binocular and monocular vision were compared. In this experiment, only one adaptation brightness (2000 ml.) was used. An advantage was found for binocular vision at seven test field brightnesses and contrast ratios. The binocular advantage varied from but a few seconds up to about 50 percent, depending upon the condition used. Although such differences might possibly be due to a summation effect, a more probable reason can be found in the relative ease of binocular viewing under the conditions used. Observing monocularly required keeping one eye closed for rather long periods of time, and such a task can become very annoying. It is quite likely that fixation

and attention could not be maintained nearly as well under such conditions.

The disparity between monocular and binocular viewing does not, of course, minimize the value of the present study for practical purposes. It merely means that some constant correction might have to be made in the present data when they are applied to the practical radar situation. Inasmuch as the correction factor probably is much less than individual differences and experimental error, any practical conclusions from the experiment would scarcely be altered.

CONCLUSIONS

In general agreement with work of other investigators employing very different test stimuli, the lowest contrast thresholds and the shortest detection times occurred when the test and adapting illuminations were equal or very nearly so. Differential brightness sensitivity was found to be considerably less when measured at an illumination differing greatly from that to which the eye had been adapted. Within the range of ordinary room illumination, the level of adaptation of the eye is an insignificant factor in radar detection.

It follows that when bright screens are used, considerably higher pre-exposure brightnesses can be tolerated, but dark adaptation may cause some loss in sensitivity. Very strong signals can be detected almost immediately even when the eye has been exposed to a relatively high brightness. The best operating levels (at least so far as adaptation conditions are concerned) have been shown to be the brighter screens; at these levels the effects mentioned are minimized.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 38, NUMBER 4 APRIL, 1948

An Investigation of the Use of Geiger-Müller Photoelectron Counters in the Spectrographic Analysis of Phosphorus in Steel

RICHARD HANAU*
Department of Engineering Research, University of Michigan, Ann Arbor, Michigan
AND
RALPH A. WOLFE
Department of Physics and Department of Engineering Research, University of Michigan, Ann Arbor, Michigan
(Received October 9, 1947)

For routine quantitative spectrographic analysis of phosphorus, using the line at 2136A, the Geiger-Müller photoelectron counter has certain advantages over both the photographic plate and the electron multiplier photo-tube. The construction of the counters and the methods of electrically shielding and mounting them on the Littrow spectrograph are briefly described. The adjustments of counter orientation, slit widths, and focus are discussed with reference to the analyses of ferrous alloys where good resolution is necessary. Because of the absorption of radiation by quartz at low wave-lengths, both absolute and relative intensities are dependent upon the position of the arc on the electrode surface. This effect, which results in appreciable error in the measured percentage composition of a sample, is investigated photographically and with G-M counters, and several methods of reducing the effect are discussed. The effect of temperature changes of the dispersing prism on line position is investigated and a suitable method of control is mentioned. Preliminary consideration of this direct-reading method shows that it seems possible to reduce the various errors so that the source fluctuation is the controlling factor.

INTRODUCTION

THE work described here was done for the Ford Motor Company of Dearborn, Michigan, in the Department of Physics under the auspices of the Department of Engineering Research at the University of Michigan. This article deals with the development of a method suitable for routine quantitative spectrographic

*Now at the University of Kentucky, Lexington, Kentucky.

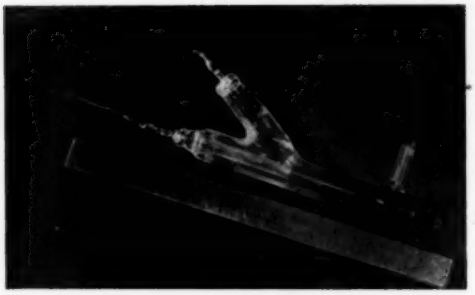


FIG. 1. The Geiger-Müller counter. The pumping stem and the side arm holding the cathode are in the same plane, allowing the counter boxes to be small and, therefore, a close line pair to be used.

analysis of phosphorus in ferrous alloys. A future paper will consider this method under industrial conditions.

In analyzing for small amounts of phosphorus with a high voltage a.c. arc, sensitivity considerations require that one of the lines 2136A or 2149A be used. For higher percentage composition, where the lines in the range 2534 to 2555A could be used, difficulties in resolution from strong iron lines will be encountered. The lines in the range 2023–2034A, though free from iron lines, can only be used at still higher percentages, particularly with prism instruments. In the present investigation, P I 2136.199A was selected as the line having the least resolution difficulties.

The use of direct-reading receivers in spectrographic analysis, such as electron multiplier tubes[1–6] and Geiger-Müller counters,[7] has the advantage of eliminating entirely the calibration, processing, and microphotometric procedures. The multiplier tube, as now available, begins to fall off in sensitivity below 2300A, while a pure quartz envelope allows the G-M counter to be used to any low wave-length accessible to a prism spectrograph.

When comparing the photographic plate, multiplier tube, and Geiger counter with regard to sensitivity, it must be remembered that the processes involved in the three receivers are different. For both the multiplier tube and photographic plate, unless the signal is above a certain minimum, it will be entirely unobserved for any length exposure. For the Geiger counter, on the other hand, there exists no absolute lower limit for a detectable signal; the lower the signal the longer the exposure necessary to separate it from noise. For a very low signal, therefore, the Geiger counter has a fundamental advantage over the photographic plate and the electron multiplier tube. However, for higher intensities, the accumulating action of the photographic emulsion is important in increasing the signal to noise ratio. In the present work a direct-reading receiver was substituted for the photographic plate on the basis of time rather than sensitivity considerations. Counters were used because long exposures could increase the signal-to-noise ratio to the point where the sensitivity was adequate. At higher wave-lengths, however, where multiplier tubes are sensitive, it may be profitable to use them as a time-saving device, even though higher sensitivity could still be attained with long exposures with Geiger-Müller counters.

APPARATUS

Most of the experimental work was done using a 2200-volt, 2-ampere, a.c. arc.[8] Preliminary work showed that a 220-volt, d.c. arc was not satisfactory for quantitative work. For analyses where sensitivity is not of prime importance, use of a spark source should be feasible. The counters and recording circuit are shielded so that, even when they are close to the inductance of a powerful a.c. spark,[9] no high frequency interference is noticeable.

The photoelectron counters, having nickel cathodes and filled with pure hydrogen, were constructed and processed as described by

[1] E. A. Boettner and G. P. Brewington, J. Opt. Soc. Am. **34**, 6 (1944).
[2] M. F. Hasler and H. W. Dietert, J. Opt. Soc. Am. **34**, 751 (1944).
[3] G. A. Nahstoll and F. R. Bryan, J. Opt. Soc. Am. **35**, 646 (1945).
[4] D. L. Saunderson, V. J. Caldecourt, and E. W. Peterson, J. Opt. Soc. Am. **35**, 681 (1945).
[5] G. H. Dieke and H. M. Crosswhite, J. Opt. Soc. Am. **36**, 192 (1946).
[6] K. G. Kessler and R. A. Wolfe, J. Opt. Soc. Am. **37**, 133 (1947).
[7] O. S. Duffendack and W. E. Morris, J. Opt. Soc. Am. **32**, 8 (1942). Our attention has been called to a recent article by Borison and Fogel, J. Exper. and Theoret. Phys., U.S.S.R. **17**, No. 5 (May, 1947), on the use of Geiger-Müller counters in quantitative spectrographic analysis.
[8] O. S. Duffendack and K. Thomson, Proc. Am. Soc. Test. Mat. **36**, 301 (1936).
[9] J. H. Enns and R. A. Wolfe, Paper presented at the A.S.T.M. Symposium on Spectroscopic Light Sources, June 20, 1946, at Buffalo, New York.

Duffendack.[10,11] Since it is sometimes desirable to have the counters unshielded and the optics exposed, it was originally thought that pure nickel cathodes would be preferable to those of alloy nickel so that the counters would be insensitive to ordinary light sources. However, no reproducible differences in light or electrical characteristics were apparent in a comparison of counters having pure and alloy cathodes.

Because it is very important, as will be indicated below, to use a line pair as close together as possible, counters were constructed with cathodes of different diameters to determine whether exceptionally small counters are practical. It was found that the length of the plateau varies approximately linearly with the cross-sectional area of the hollow cathode. More extensive experiments by Christoph[12] show that the plateau of the count-voltage curve is greatly decreased if the anode wire is asymmetric with respect to the cathode. Because of this effect, which places a lower limit on the size of the counter, we believe that future work should use deflecting mirrors[13] in order to utilize a closer line pair.

The counters (see Fig. 1) have nickel alloy cathodes 6.75-mm O.D. and 5.95-mm I.D. The cylinder is $1\frac{1}{4}''$ long with a slot $\frac{1}{8}''\times\frac{5}{16}''$ cut lengthwise in the side wall to receive the radiation. The anode is a tungsten wire 3 mils in diameter. The envelope of the counter is pure quartz and contains hydrogen at a pressure of 90 mm. The counters have threshold voltages of about 700 volts, plateaus of about 150 volts, and a background count of about 5 per minute. Each counter has a scale-of-eight recording circuit, similar to that described by Lifschutz.[14] The mechanical register, which limits the resolving time of the entire circuit, is capable of registering 3600 counts per minute.

The counters are used with a Bausch & Lomb Littrow spectrograph, as shown in Fig. 2. To afford rapid interchangeability between the direct-reading and the photographic methods, a Duralumin plate with a central, longitudinal slot slides into the spectrograph camera in the same manner as the plate holder. The Duralumin plate supports a fixed screw and two brass rails. The brass counter boxes move on the rails by means of a large nut turning on the fixed screw; the position of a counter may be read to within 1μ.

The counter boxes, each completely enclosing a counter, are necessary to shield against high frequency disturbance. Two shielded connectors at the top end of each box connect the counter to the shielded leads going to the Neher-Harper quenching circuit.[15] The outside shield on the cable from the quenching circuit to the scaling circuit is flared and electrically attached to the outside of the metal cabinet. The entire receiver, therefore, consisting of counters, leads, quenching circuit, cable, and scaling circuit is completely enclosed in a single continuous metal shield.

The counter slits are of stainless steel, lapped to an angle of 20°, attached to the outside of the counter boxes. It is possible to make adjustments on these slits by working through the slot cut in the Duralumin plate, thus eliminating the necessity of unfastening a counter box from the mounting. Because of the inability of exactly setting the counters at the same position relative to a line and because of temperature shift of lines

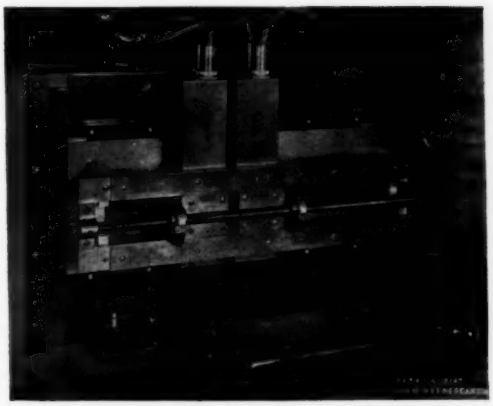


FIG. 2. The two Geiger-Müller counters mounted on the Bausch & Lomb Littrow spectrograph. The leads to the quenching circuits come out at the top of the counter boxes. The entire mounting is removed as a unit from the camera by sliding it horizontally to the right as in the case of the photographic plateholder.

[10] O. S. Duffendack, H. Lifschutz, and M. M. Slawsky, Phys. Rev. **52**, 1231 (1937).
[11] O. S. Duffendack and W. E. Morris, J. Opt. Soc. Am. 32, 8 (1942).
[12] W. Christoph, Ann. d. Physik **26**, 145 (1936).
[13] J. L. Saunderson, V. J. Caldecourt, and E. W. Petersen, J. Opt. Soc. Am. **35**, 681 (1945).
[14] H. Lifschutz, Rev. Sci. Inst. **10**, 21 (1939).
[15] H. V. Neher and W. W. Harper, Phys. Rev. **49**, 940 (1936).

discussed below, a counter slit comparable in width to the spectrograph slit is impractical. Because of low intensities, and because a flat-topped line contour is not always a good approximation to the true intensity distribution, a counter slit much narrower than the line is also impractical. In view of these considerations, the counter slits are set at 200μ. Since the spectrograph slit is about 30μ, the counter slit is about three times the width of a spectral line along the focal plane, which agrees with the figure used by Hasler.[16]

In a Littrow instrument, the angle that the incident light makes with the plane of the counter slit varies because of three factors, namely, (a) the change of plate-holder tilt, (b) the change of counter position, and (c) the change of incident light direction due to arc wandering. Factor (a) results in a total change in angle of about 6°. The magnitude and sign of the variation due to (b) depends on whether the counter moves along the focal curve, or along a chord to this curve, but is of the order of 2°. Factor (c) results in a change of about 1 to 2°, depending on the size of the electrode image at the prism. A versatile receiver, therefore, capable of being used from 2000 to 8000A, must be able to receive incident light when the total variation in angle is more than 10°.

A central aim throughout this investigation has been to adapt the spectrograph for direct reading in such a way that the photographic plate may be readily and rapidly used at any time. As regards focusing, this means that the plane of the counter slits must be brought to the position normally occupied by the photographic emulsion. The mounting for the G-M counter is so constructed that by loosening four screws various spacers and shims can be inserted to bring the plane of the slits to any desired position. The spacers are of such thickness that the counters (and therefore their slits) can be moved through a distance of $\frac{1}{4}$" in steps of 0.009" or less, which is less than can be noticed in determining the position of best focus. The correct spacer thickness for best focus may be found by taking line contours as a function of position of the counter with a narrow counter slit. Graphs of maximum intensity or half-width plotted against different thickness of spacers will show the best value. A faster method is the insertion of a small piece of photographic plate so that the emulsion is in the plane of the counter slit, the counter having been moved aside.

The spectrograph slit and counter slit are brought into parallelism by the above method of contours. This adjustment is an additional "degree of freedom" which must be considered in using direct-reading receivers where good resolution is needed.

One way of setting the counters is to run over the line listening to the counts. Our investigation has shown that this method is only satisfactory with very steady sources. For accurate work the two counters should be set in unison by using the intensity ratio. After setting one counter approximately by ear, the intensity ratio is found for various settings of the other. The extremum

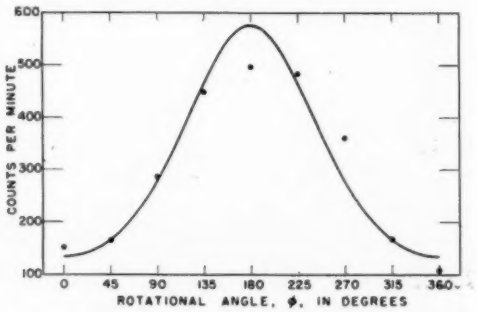


FIG. 3. Intensity as a function of rotation angle, φ. The full curve is the equation $I=278\exp(-0.729\cos\varphi)$ and the dots are experimental points using the copper line 2135.976A. The angle φ measures the position of the arc on the electrode circumference and is zero when the light passes through a maximum thickness of quartz.

[16] M. F. Hasler and H. W. Dietert, J. Opt. Soc. Am. **34**, 751 (1944).

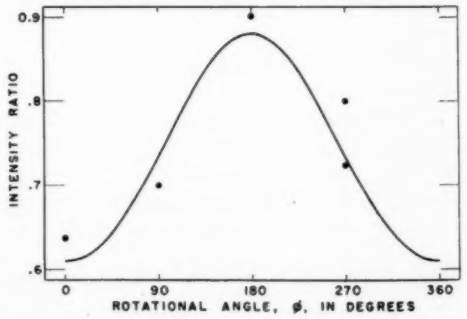


FIG. 4. Intensity ratio as a function of rotation angle, φ. The full curve is the equation $R=0.734\exp(-0.184\cos\varphi)$ and the dots the experimental values of the ratio of the two copper lines, 2135.976 to 2276.253A.

of this curve is the setting for the counter that was moved. The other counter is now set in the same way. In the present investigation, after careful focusing, slit aligning, and setting, it has been found possible to measure the total intensity of the weak phosphorous line, 2136.199A, without recording any intensity due to the very strong copper line, 2135.976A, when the spectrograph slit is 35μ. On the spectrograph used for most of the experimental work, the centers of these lines are separated by less than 150μ, measured along the focal curve.

EXPERIMENTAL RESULTS

The Effect of Quartz Absorption

Because the arc wanders erratically over the ends of the electrodes, the light passes through varying thicknesses of quartz, caused principally by the varying light path in the dispersing prism. If the electrodes are cut so as to leave a small projection at the circumference, the arc may be localized to this small area by opposing the two projections. By rotating the electrodes about their axes through an angle, φ, the light passes through different thicknesses of quartz. If I is the emergent intensity,

$$I = C_1 \exp[-C_2 \cos\varphi], \qquad (1)$$

where C_1 and C_2 are constants for a given wave-length and for a given electrode at a fixed position. Figure 3 shows a curve of this equation and some experimental points for three-minute exposures. The deviation from theory can be accounted for by the experimental error in the measurement of φ and the statistical error in the counting rate.

It is seen that the intensity, as measured by a single counter (Fig. 3), changes by a factor of four as the arc wanders about the electrode. This variation is greatly reduced, but not eliminated, by using a reference line and measuring the ratio of intensities. If two lines, λ_1 and λ_2, are measured and k_1 and k_2 are the corresponding absorption coefficients, then the intensity ratio is

$$R = (I_{01}/I_{02}) \exp -[(k_1 - k_2)x], \qquad (2)$$

where x is the thickness of quartz traversed. Since k is a function of wave-length, the ratio is a function of thickness of quartz traversed. Equation (2) may be put into the same form as (1). Figure 4 shows this curve and experimental points for the ratio of two copper lines.

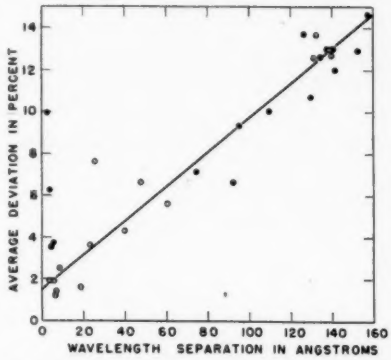


FIG. 5. Average deviation of a single determination of intensity ratio from the mean value as a function of wave-length separation of the two lines, measured photographically. Dots represent deviations of ratios calculated with line P 2136.199A, open circles with line Fe 2270.860A. The intercept at zero wave-length separation is considered to represent the emulsion and source errors. The fluctuation of the points about the curve is due to the fact that the intensity ratio is a function of mean wave-length as well as wave-length difference.

In the above discussion, we have considered that the change in intensity is due to the dispersing prism only. Actually, both the foreprism and the collimating lens will have some effect on intensity, as the arc moves on the electrode. (In these experiments no condensing lens was used.) As the arc wanders to either side, the distance traversed in the foreprism increases as secθ (where θ is the angle of refraction at the first face) while the distance traversed in the lens decreases. Because these effects are small, they have been neglected in comparison with the effect of the dispersing prism. If the losses due to reflections at the dispersing prism are considered, the results are unchanged, since the reflection coefficient, though dependent on wave-length, is independent of φ.

It is evident from Eq. (2) that the ratio of two lines depends on the position of the arc on the electrodes, only because k is a function of wave-length. We should expect, therefore, a larger dependence on arc position for widely separated lines than for lines close together. This effect was measured by recording photographically a series of equal exposures under the same source conditions. Since the arc wanders over the flat surface

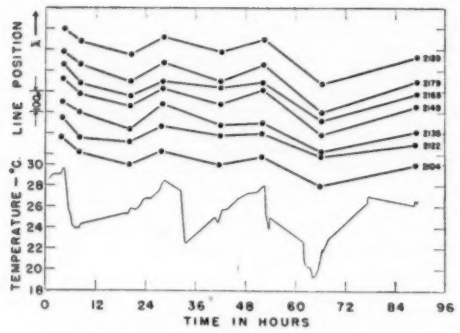


FIG. 6. Line position as a function of time for seven copper lines in the region 2104 to 2190A. Plotted below are temperature readings taken at the dispersing prism. An increase in temperature results in an apparent increase in wave-length.

of the electrodes in a somewhat random manner, a series of extremely long exposures should be identical because the arc would cover the area in a completely random manner. However, for a series of exposures comparable to the times usually used in spectrographic analysis, it is possible that the ratio of two lines may differ because the arc may not cover the area in the same way for these short exposures. Therefore, the average deviation from the mean of a single determination of intensity ratio measures the effect of arc wandering for the particular time of exposure used. Figure 5 shows this average deviation plotted against the wave-length separation of the two lines used.

The ratios were calculated between seventeen lines in the wave-length region 2130 to 2294A. For a given wave-length separation, the ratio will depend on the mean wave-length since the factor (k_1-k_2) is a function of wave-length as well as wave-length difference. It is possible to form ratios involving different mean wave-lengths for approximately the same wave-length difference. Since the factor (k_1-k_2) will be different for these cases and therefore the effect of arc wandering will be different, we should expect the average deviation to depend on mean wave-length as well as wave-length separation. This effect probably accounts in part for the scattering of the data.

The error caused by quartz absorption may be reduced by making the arc wander more erratically over the surface, thereby covering the whole area randomly in a finite exposure. An a.c. spark or an arc which is spark-like in character should prove helpful.

A second method of decreasing this error is to localize the arc by grinding the electrodes in the form of a cone so that the arc remains near the center. This method is useful for short exposures only, because the electrodes flatten off fairly quickly. For longer exposures the electrodes are ground in the form of a wedge, the two cutting edges parallel to each other and to the source-slit distance. The different-shaped electrodes greatly reduced the effect but did not eliminate it entirely. An easy method of testing whether the absorption effect has been successfully eliminated for any given set of conditions is to plot the intensity ratio of the two lines against the absolute value of either one. If the graph shows no correlation, the absorption effect is absent. In addition to changing the source conditions and the electrode shape, the use of a close line pair is very necessary, both in direct-reading receivers and in analyses done photographically.

The Effect of Temperature on Line Position

The effect of temperature changes on line position, though unimportant for ordinary exposures when a photographic plate is used, must be considered when employing a direct-reading receiver, where the entrance slit of the counter or photo-tube is usually small. Figure 6 shows the dependence of line position on temperature. The temperature readings are not equilibrium readings, which may explain the apparent lag in line position relative to temperature changes.

The effect of temperature may cause changes in the spectrograph, the optics, or the counter mounting, thereby changing the position of the line relative to the counter slit. However, analysis of our apparatus regarding the effect of the spectrograph slit, the spectrograph base, the mounting of the dispersing prism, and the mounting of the G-M counters shows that these cannot account for the shift, either together or separately. Sawyer[17] and Merton[18] have discussed the effect of temperature on the dispersing prism, and have

[17] R. A. Sawyer, *Experimental Spectroscopy* (Prentice-Hall, Inc., New York, 1944), p. 118 ff.
[18] T. R. Merton, Proc. Roy. Soc. **113A**, 704 (1927).

indicated methods of reducing it. If the index of refraction, n, is both a function of temperature and of wave-length, we may write

$$d\lambda/dT = (dn/dT)/(dn/d\lambda). \qquad (3)$$

For the mean wave-length of the lines investigated, $d\lambda/dT$ is positive as found from Eq. (3) which agrees with Fig. 6. The available values of n as a function of T and λ gives a shift of wave-length somewhat lower than is observed.

Since the shift in line position is several times the error in setting the counters, it is necessary to eliminate this effect. Resetting the counters at frequent intervals is impractical in routine work. An empirical relation between counter settings and temperature could be established, but this would suppose that the prism is in equilibrium at the temperature of the thermometer. A more satisfactory method, which is being used at present, is to completely enclose the spectrograph in an insulated box, controlling the temperature to better than ± 0.5°C by means of a mercury column thermostat. Present results show that any variation in line position is smaller than the error in determining this position.

DISCUSSION AND SUMMARY

The use of Geiger-Müller counters for spectrographic analyses seems entirely possible. Besides the usual precautions necessary when using the photographic method, certain additional conditions must be considered, such as temperature control and parallelism of the slits. The selection of a reference line close to the analysis line, desirable at all wave-lengths in photographic work because of calibration considerations, is particularly necessary in regions where the spectrograph optics absorb appreciably. By selection of a very close line pair it seems possible to reduce this error below the emulsion error. For direct-reading instruments, the use of deflecting mirrors to utilize close line pairs seems entirely feasible, and by this means the error due to arc wandering can probably be made less than the statistical error.

Non-self-quenching counters, of the type described here, appear to be stable in operating characteristics. With plateaus of the length usually attainable, tests have shown that frequent adjustment of the circuit controls introduces errors smaller than the expected statistical errors. The determinations of exposure time and time of wait are made by considering the magnitude of the expected statistical error. Exposure times must be long enough to reduce this error so that it may be neglected relative to source and other errors. Time of wait should be such that, increasing it has less effect on the reproducibility of the intensity ratio than is caused by statistical variation.

A future paper will consider in detail the stability of the analytical curves and reproducibility of individual determinations. Preliminary results indicate that the intensity ratio is sensitive to small variations in gap setting, electrode length, and electrode diameter. It appears, however, that, just as in the case of the photographic method,[19,20] the many individual errors can be reduced to such a point that the source error is the determining factor.

[19] R. M. Sawyer and H. B. Vincent, J. Opt. Soc. Am. **31**, 47 (1941).

[20] H. H. Grossman, R. H. Sawyer, and H. B. Vincent, J. Opt. Soc. Am. **33**, 185 (1943).

Effects of Illumination Level on Visual Performance and Fatigue*

ERNST SIMONSON AND JOSEF BROZEK**
The Laboratory of Physiological Hygiene, University of Minnesota, Minneapolis, Minnesota
(Received October 14, 1947)

The purpose of this study was to investigate the relation of the level of illumination intensity to performance and fatigue in visual work. The question of illumination requirements is of considerable importance from the point of view of physiological optics, industrial physiology, and public health. Yet a large part of the research done in the past leaves much to be desired in regard to general approach and specific testing techniques as well as experimental analysis and statistical evaluation. In the present series six illumination levels (2, 5, 15, 50, 100, and 300 footcandles) were studied in repeated experiments in six normal subjects. The work task involved recognition of fine details (letters) and reproduced the essential features of a conveyor inspection operation. In addition to the studying of various criteria of work performance and their change in the course of 2 hours of work, a large battery of visual tests was applied before and after the work so as to characterize the degree of fatigue. Varied illumination affected the performance more than the functional criteria of fatigue. Only one function, the recognition time for stimuli of threshold size, showed a decreasing degree of fatigue up to 300 ft.-c, all other variables, including performance, which changed with a changing level of illumination showed an optimum at 100 ft.-c. The demonstration of an optimum level of illumination makes highly questionable the current practice of recommending "minimum" levels of illumination for industrial jobs, since deterioration of performance and increase in fatigue may result when the optimal level is exceeded. The *optimum* of 100 ft.-c for the present strenuous visual task is at or below the *minimum* which should be recommended on the basis of the code of the Illuminating Engineering Society.

INTRODUCTION

THE importance of the intensity of illumination for visual work has been recognized for a long time and has found practical expression in illumination recommendations for the different categories of jobs. During the past 35 years the recommended illumination intensity has been steadily increased. Is there a valid basis for this trend?

The illumination code of the Illuminating Engineering Society (I.E.S.), recommending minimum levels of illumination,[1] has been generally accepted in this country.[2,3] It is claimed that adherance to these recommendations would contribute to conservation of eyesight, attraction of a better class of employees, decrease of the accident rate, improvement of quality and quantity of production and of the morale of employees.[4,5] Little actual evidence is available to substantiate these claims. Studies on which the I.E.S. recommendations were based,[4–11] including the effect of illumination on output in industrial work, largely fail to stand the test of a critical analysis.

In the recommendations of the I.E.S., much reliance is placed also on the experimental work of Luckiesh and Moss[12] which has been repeatedly and severely criticized. These authors recommend as much as 500 to 3000 footcandles (ft.-c) for

* Aided by a research grant from the Verd-A-Ray Corporation, Toledo, Ohio.
** With the assistance of Dr. W. J. Bushard and Dr. H. G. Peterson of the Division of Ophthalmology and Dr. J. C. Franklin and Mr. A. Butler of the Laboratory of Physiological Hygiene.

[1] "American recommended practice of industrial lighting," Illuminating Engineering Society (1942).

[2] J. S. Rogers, Illumination. In "Protecting Plant Manpower," pp. 63–66, U. S. Dept. of Labor, Washington, 1941.

[3] A. D. Brandt and H. E. Seifert, "Illumination, noise and radiant energy," *Manual of Industrial Hygiene* (W. M. Gafafer) (W. B. Saunders, Philadelphia, 1943), pp. 353–372.

[4] Committee on Industrial and School Lighting, "Report on lighting in the textile industry," Trans. Illum. Eng. Soc. **32**, 247 (1937).

[5] Committee on Industrial and School Lighting, "Report on lighting in the shoe manufacturing company," Trans. Illum. Eng. Soc. **32**, 289 (1937).

[6] Committee on Industrial and School Lighting, "Progress report on the lighting in the printing industry," Trans. Illum. Eng. Soc. **31**, 277 (1936).

[7] Committee on Industrial and School Lighting, "Report on lighting in the candy manufacturing industry," Trans. Illum. Eng. Soc. **32**, 483 (1937).

[8] Committee on Industrial and School Lighting, "Studies in lighting of intricate production, assembly and inspection processes," Trans. Illum. Eng. Soc. **32**, 1019 (1937).

[9] Committee on Industrial and School Lighting, "Lighting for silk and rayon throwing and wide goods weaving," Trans. Illum. Eng. Soc. **33**, 17 (1938).

[10] Committee on Industrial and School Lighting, "Lighting for the machining of small metal parts," Trans. Illum. Eng. Soc. **34**, 21 (1939).

[11] Committee on Industrial and School Lighting, "Lighting of power presses," Trans. Illum. Eng. Soc. **34**, 153 (1939).

[12] M. Luckiesh and F. M. Moss, *The Science of Seeing* (D. Van Nostrand Company, Inc., New York, 1938), p. 548.

certain jobs requiring discrimination of fine details. In general, for such types of visual work, 100 ft.-c or more are given as *minimum* requirements.[1] It is suggested that higher illumination levels than those recommended are desirable, or at least acceptable. The optimal levels are not stated, but from Luckiesh and Moss'[12] discussion it may be inferred that the desirable levels exceed the recommended levels by at least two or three times, and no hint has been given that high illumination levels might produce a deterioration of performance, even when glare is excluded. Such an approach appears to be totally unphysiological. It has been shown for other types of environmental factors that an optimum of performance or efficiency is attained at a certain medium value of the independent variable. This is very generally the rule; one wonders why a different law should apply to illumination intensity.

The first requirement for research in this area is the development of suitable work tasks. These should meet the following requirements: (1) easy standardization of work performance, (2) possibility of quantitative evaluation, (3) clear-cut relationship to fundamental visual functions, (4) elimination of auxiliary functions such as manual skill or verbal intelligence which might influence the results significantly, (5) negligible effects of practice, (6) practical applicability of results, and (7) possibility of varying critical factors of performance. The work task developed in this laboratory for the study of visual performance reproduces the job of close inspection of small objects transported on a conveyor, and appears to meet the above requirements.[13]

Both maximum performance and the capacity for a sustained effort are important. In the past, laboratory experiments on vision as a rule have been very short in duration. Thus in some of Weston's work[14] the test time was only one minute. The visual work test developed here at the laboratory can be continued, theoretically, for an indefinite period of time. In practice we found that a two-hour spell of uninterrupted work was about as much as trained subjects would stand. This is long enough to provide information about the initial effect of illumination on performance as well as on the changes resulting from continued work.

In view of the frequency of complaints about visual fatigue, there is surprisingly little information available about the fatigue trends of various fundamental visual functions accompanying visual effort. Therefore, in addition to the study of performance, attention was paid to functional changes. A large battery of tests was administered immediately before and after the work. There is evidence[15,16] that the results obtained for one visual function may not apply to other functions. Therefore, any conclusions and recommendations should be based on an integrative interpretation of multiple visual functions and characteristics, including performance.

METHODS

The Work Test

The work test has been described in detail elsewhere.[13] The visual task consisted in the recognition of small letters (size corresponding to a visual angle of 10 minutes) which were presented in random order behind a narrow slit. The head was kept at a constant distance from the screen by means of a forehead rest. The letters, printed on dull white paper, were mounted on a rubber belt driven by a synchronous motor. The exposure time from the first appearance of the letter in the slit to complete disappearance was 0.56 sec. The subjects copied down the letters by hand. The writing of the letters was done without visual control, so that the eyes could remain focused at the screen throughout the work spell. The writing of letters did not require any appreciable manual skill.

The lamps were mounted in boxes shielded from the subject, and the illumination level was adjusted primarily by means of interchangeable perforated diaphragms placed in front of the lamps, and finally by varying the distance of the

[13] J. Brozek, E. Simonson, and A. Keys, "A work test for quantitative study of visual performance and fatigue," J. App. Psychol. **31**, 519 (1947).

[14] H. C. Weston, "The relation between illumination and industrial efficiency: I. The effect of the size of work." Joint Rep. Industr. Health Res. Board and Illum. Res. Comm. (London) (1935).

[15] F. N. Freeman, "An investigation of the illumination requirements of school buildings," National Education Association: Report of Committee on School House Planning, 90 (1925).

[16] H. J. Ströer, "Rationalisierung der Arbeitsplatzbeleuchtung," Industr. Psychotechnik **3**, 289 (1926).

lamps from the slit. The light distribution on the surrounding field was essentially identical at the various illumination levels investigated. By using bulbs of 40, 100, and 200 watts, a range from 2 to 300 footcandles (ft.-c) was obtained. The six subjects were seated in individual booths.

The standard test involved 2 hours of continuous visual work. The performance was evaluated in terms of the number of letters correctly recognized out of a sample of 200 letters; this corresponded to a time interval of 6 minutes. Three such work samples were evaluated. The first sample was taken beginning 5 minutes after the start of the work, a second sample at the midpoint (after 1 hour of work), and the third sample near the end after about 110 minutes of work. The subjects had the impression that the whole 2-hour performance was evaluated, but this was actually done in only three experiments; it was shown that the average of the three samples was a fair representation of the total performance.

The average of the three samples gives the "average performance score." In addition, the difference between the first and the third sample ("performance drop") and the difference between the highest and lowest value of the three samples ("performance range") was calculated. The latter is an index of the maximal intra-individual variability during a single work period. The "performance drop" served as an index of fatigue.

Blinking Rate During Work

The blinking rate was measured at the start, midpoint, and end of work. The blinks were counted for periods of 5 minutes timed by means of a stop watch. The results were expressed as "average blinking rate per minute" (means of the 3 periods) and "blinking rate difference" (change from the first to the last period).

Before and After Tests

Immediately before and after the work a battery of 4 groups of visual tests was applied, each directed by a different experimenter and spaced at intervals of about 5 minutes. The subjects were spaced in time accordingly, starting (and finishing) their 2-hour work task at 5-minute intervals. The 4 groups of visual tests were:

I. Recognition time for dots of threshold size, and fusion frequency of flicker (f.f.f.).

II. Abduction power, adduction power, vertical divergence, accommodation near point, and convergence near point; except for the convergence near point, which involved both eyes, all of these measurements were made on the dominant eye.

III. Ophthalmograph (9 items).

IV. Brightness discrimination (red and green).

After completing the tests, the subjects were asked to fill out a questionnaire. A brief description of the methods will be given.

Recognition Time for Stimuli of Threshold Size

Visual acuity was determined by means of small black dots of varying size, on white paper squares pasted on a disk. The squares with the dots were presented through a window, 1.5 by 2.5 inches. The subjects had to locate the dot; the placement varied in the different squares. The time from the start of the exposure to recognition was measured by means of a stopwatch, and the value for the smallest dot which could be seen was used as the "recognition time." The illumination of the test patch was constant.

Fusion Frequency of Flicker (f.f.f.)

The rate of flashes was automatically increased from coarse flicker until the test patch appeared plain. The rate of flashes was read on a scale. The moment of fusion was indicated by the subject through an acoustic signal.

Abduction Power, Adduction Power, and Vertical Divergence

The measurements were made by means of the Risley rotating prism, held in front of one eye, while the subject was fixating with both eyes a spot of light 6 meters distant. The prism strength was increased until diplopia occurred. Each measurement was obtained in triplicate, and the average was recorded. The readings are expressed in prism diopters.

Accommodation Near Point

The accommodation near point was determined by the Prince Rule, manufactured by the American Optical Company. The rule was supported on

the infra-orbital bony rim. A card with small letters, corresponding in size to the "14/21" line in acuity tests for near vision, was brought toward the eye from the range of clear vision to the point at which the letters started to become blurred. The reading was recorded in terms of the distance of point of blurring from the bony rim, in cm. Three consecutive measurements were made and averaged.

Convergence Near Point

The Prince Rule was held against the bridge of the nose, and a white circular target, 2 mm in diameter, was moved from the distance of clear vision toward the eye to the point of diplopia. The distance from the bridge of the nose to the point of diplopia was measured in millimeters. Three consecutive readings were made and averaged.

Ophthalmograph

By means of a beam of light reflected from the cornea, the movement of both eyes were recorded on a moving film. The subject was asked to change the fixation between two targets, printed on the left and the right side of a white test card, as fast as possible. Eye movements were photographed for a period of 10 seconds. The curves, recorded on transparent film, were projected onto a calibrated screen for the purpose of evaluation. Nine different characteristics were measured, but only 6 were analyzed in this series: "Movement phase" refers to the time interval of movement between two fixation points. "Angular velocity" is obtained by dividing the "movement phase" through the "distance of movement," expressed in arbitrary units. "No. of movements" expresses the rate of movement per sec. "Fixation phase" is the time of the cycle used for fixation of the targets. Consequently, the total cycle length equals the movement phase plus the fixation phase. The "extent of movement during fixation" is the amplitude of motion during the fixation phase, serving the purpose of a more accurate fixation; the eyes were frequently arrested near the target with some over- or under-shooting so that some adjustment was still made during the fixation phase. This does not mean that the final adjustment is necessarily accurate; the difference between actual distance of the two points of fixation and the true distance of the targets is referred to as "the discrepancy between fixation point and target." The technique will be described in greater detail elsewhere.[17]

Brightness Discrimination

The Pulfrich photometer (Zeiss) was used for this test. The left half of the visual field was set at a standard brightness level. The right half of the visual field was set at zero level (darkness), and the subject was asked to increase the brightness, by turning a knob, until both halves matched. Whenever the subject reported that he exceeded this point, no correction was permitted but the reading was repeated. The accuracy of matching was determined in arbitrary units as the difference between the true brightness of the right and the left halves. Five consecutive measurements were made and averaged. The brightness discrimination was investigated first with green and then with red filters.

The sequence of the tests, although arbitrary, was kept unchanged through the whole series. The selection of tests was made in order to obtain information about a maximum number of visual functions within 20 minutes. It was thought that it might be difficult to discover fatigue trends persisting for more than 20 minutes after the end of the work task.

After completing the brightness discrimination test, the subjects characterized their subjective responses in each experimental session by means of a standardized questionnaire containing the following 10 items: Smarting sensation in eyes, watering of eyes, feeling of strain in eye muscles, difficulty in focusing (blurring of letters), tendency to blink, difficulty in covering the whole height of the slit, difficulty in maintaining continuous attention, time exposure of letters appearing shorter, headache, and over-all feeling of tiredness. The items were rated on a 5-point scale, ranging from "0" (absent) to "4" (extreme), and the score was calculated as the added sum of the 10 items.

In some of the functions measured an increase in the score would indicate deterioration (for in-

[17] J. Brozek, E. Simonson, and A. Keys, "Quantitative analysis of voluntary eye movements," (to be published).

stance, recognition time for threshold stimuli, accommodation near point), in others, an improvement (work output, flicker fusion frequency). To allow uniform presentation of the material and a better visualization of the trend of the results, improvement in all functions is indicated by "+," and deterioration by "−." Table I shows what direction of change in the original readings would correspond to an improvement, i.e., could be classified as "desirable" change.

In all functions, except average performance score and average blinking rate, we are concerned with the differences (d) in the scores obtained before and after the visual task, or at the beginning and at the end. Therefore, for the purpose of condensation, only the mean differences ($\bar{d}$), and not the absolute figures, are given in the tables. The differences between the $\bar{d}$ at different illumination levels are designated as $\bar{\Delta}$.

SUBJECTS

The subjects were six young men (ages 20 to 29) who had served in the laboratory for the year preceding the visual research as full-time experimental subjects in nutritional investigations. They were physically normal, intelligent, and cooperative, well accustomed and adjusted to the routine of laboratory work and testing. Their diet and daily routine and work was known, and any major interference from personal habits or accidental stresses could be excluded for the three months of visual experiments. Before the start of these studies, the visual characteristics of the subjects were thoroughly investigated and found to be within normal limits. This examination was repeated at the end of the testing period. The individual ophthalmological data were given elsewhere.[18]

TABLE I. Units of functions and the "desirable" values of the scores.

Retinal functions	Units	"Desirable" values
Recognition time	Seconds	Lower
Flicker fusion frequency	Flashes per second	Higher
Brightness discrimination, green	Arbitrary units	Lower
Brightness discrimination, red	Arbitrary units	Lower
Ophthalmological Tests		
Abduction	Diopters	Higher
Adduction	Diopters	Higher
Vertical divergence	Diopters	Higher
Accommodation near point	Centimeters	Lower
Convergence near point	Centimeters	Lower
Performance Criteria		
Performance average	No. of correct letters	Higher
Performance drop	No. of correct letters	Lower
Performance range	No. of correct letters	Lower
Blinking Rate		
Blinking rate average	Blinks per minute	Lower
Blinking rate difference start-finish	Blinks per minute	Lower
Questionnaire Score	Arbitrary units	Lower
Ophthalmographic Measurements		
Movement phase	1/100 sec.	Lower
Angular velocity	Arbitrary units	Higher
No. of movement	Rate per second	Higher
Fixation phase	1/100 sec.	Lower
Extent of movement	Arbitrary units	Lower
Discrepancy	Arbitrary units	Lower

CONDITIONS

Three different commercially available illuminants (usual frosted lamps, natural-white lamps, and Verd-A-Ray lamps) were used at each level of illumination. Differences in visual performance and associated changes of visual functions between the illuminants will be discussed in another communication, together with data on spectral distribution. In this paper the data obtained for different illuminants at the same level of illumination intensity will be combined. On each experimental day all 3 illuminants were used, 2 subjects working with each lamp. The subject pairs and illuminants were rotated for consecutive experiments. Each combination of subject pairs and illuminants was repeated twice.

The three types of lamps were investigated at 5, 100, and 300 ft.-c. The mean scores, obtained at each of these levels of illumination, are based on 36 observations. In addition, three other levels (2, 15, and 50 ft.-c) were investigated, but only with one or two illuminants and a smaller number of observations (6 or 12). The experiments were performed in an air-conditioned suite (77°F, 50 percent relative saturation).

The subjects were completely disinterested in the specific results of these studies and were not aware of the specific factors investigated in a particular testing session since they worked in individual booths and the lamps were concealed. Additional precautions were taken to eliminate factors which have marred the results of many of the previous studies. The subjects were trained in the experimental procedures before the start of the experimental work proper. A trend analysis was made to show that they had arrived at a

[18] J. Brozek, E. Simonson, W. J. Bushard, and J. H. Peterson, "Effects of practice and the consistency of repeated measurements of accommodation and vergence," Am. J. Ophthal. (in press).

TABLE II. Mean scores for performance, blinking rate, and questionnaire score at three levels of illumination, and significance of the fatigue changes taking place during 2 hours of visual work.

	Level of illumination					
	5 ft.-c		100 ft.-c		300 ft.-c	
Function	mean		mean		mean	
Performance average	167.4		189.3		188.0	
Performance range	22.6		9.2		12.9	
Blinking rate average	9.77		5.80		8.32	
	$\bar{d}$	F	$\bar{d}$	F	$\bar{d}$	F
Performance drop	−14.1	29.41**	−6.8	21.46**	−10.9	30.47**
Blinking rate difference	− 0.83	1.01	−0.48	0.89	− 0.72	0.90
Questionnaire score	− 8.4	87.61**	−7.7	62.23**	− 8.2	58.19**

$\bar{d}$ equals the mean change in 36 pairs of measurements made at the start and at the end of the visual work; negative $\bar{d}$ values express deterioration.

F expresses the statistical significance of the mean changes ($\bar{d}$).

* F at 5 percent of significance is 4.13.

** F at 1 percent of significance is 7.44.

plateau of performance when the investigations started.[18]

TECHNIQUES OF EVALUATION AND PRESENTATION

The results will be presented in a series of tables and graphs; the technique of graphical presentation used in this paper makes comparable the changes occurring in the several functions at different levels of illumination.

The statistical significance of the mean difference between values obtained under any two conditions may be calculated by means of the F-test for paired variates. If Δ denotes the difference in the scores X_1 and X_2 obtained by an individual under the two conditions of illumination

$$\Delta = X_1 - X_2, \tag{1}$$

the hypothesis that the mean difference, $\bar{\Delta}$, is not significantly different from zero may be tested by an equation

$$F = \frac{(\Sigma\Delta)^2/n}{\dfrac{\Sigma\Delta^2 - (\Sigma\Delta)^2/n}{n-1}}, \tag{2}$$

where n is the number of individual determinations. This value of F is associated with 1 and $(n-1)$ degrees of freedom.

TABLE III. Visual functions which exhibited significant trends with variation of the illumination level.

	Compared levels of illumination					
	5:100 ft.-c		5:300 ft.-c		100:300 ft.-c	
Function	$\bar{\Delta}$	F	$\bar{\Delta}$	F	$\bar{\Delta}$	F
Recognition time	− 2.6	6.79*	− 3.6	8.58**	−1.0	1.53
Performance average	−21.9	102.67**	−20.6	112.99**	1.3	1.57
Performance drop	− 7.2	8.49**	− 3.2	1.85	4.1	8.69**
Performance range	−13.4	40.72**	− 9.6	17.38**	3.8	8.56**
Blinking rate average	− 3.97	49.80**	− 1.44	3.19	2.52	15.93**

$\bar{\Delta}$ equals the mean difference between 36 pairs of changes (d) at the two compared levels of illumination.

Negative $\bar{\Delta}$'s indicate greater deterioration under the first of the conditions compared.

F expresses the statistical significance of the mean difference ($\bar{\Delta}$) between the two compared levels of illumination.

* F at 5 percent of significance is 4.13.

** F at 1 percent of significance is 7.44.

The F-value derived from formula (2) is equivalent to the square of the t-value, obtained as a ratio of the mean difference between the two conditions, $\bar{\Delta}$, to the standard deviation of the difference, S.D.$_{\bar{\Delta}}$,

$$t = \frac{\bar{\Delta}}{\text{S.D.}_{\bar{\Delta}}}. \tag{3}$$

In presenting the data graphically, the averages obtained at different levels of illumination were expressed in the form of a modified t'-ratio.[19] This t'-value was computed as

$$t' = \frac{\bar{\Delta}'}{\langle\text{S.D.}\rangle_{\bar{\Delta}}}. \tag{4}$$

The numerator refers to a difference between a mean for a given level of illumination ($\bar{X}_i$) and the grand mean ($\bar{X}$):

$$\bar{\Delta}' = \bar{X}_i - \bar{X}. \tag{5}$$

The denominator is the average of the standard deviations of the mean differences between pairs of illumination levels,

$$\langle\text{S.D.}\rangle_{\bar{\Delta}} = \frac{\sum_1^c \text{S.D.}_{\bar{\Delta}}}{c}, \tag{6}$$

[19] J. Brozek, E. Simonson, and H. Alexander, "Statistical comparison of changes in fitness characteristics," (in press).

TABLE IV. Fatigue changes in retinal functions resulting from 2 hours of visual work at three levels of illumination.

	Level of illumination					
	5 ft.-c		100 ft.-c		300 ft.-c	
Function	$\bar{d}$	F	$\bar{d}$	F	$\bar{d}$	F
Recognition time	−3.69	16.55**	−1.10	3.69	−0.12	0.02
Flicker fusion frequency	−0.93	18.33**	−0.90	17.16**	−1.15	28.68**
Brightness discrimination, green	0.05	0.15	−0.12	1.31	−0.09	1.67
Brightness discrimination, red	−0.20	2.69	−0.03	0.08	0.04	0.25

$\bar{d}$ equals the mean change in 36 pairs of measurements made before and after visual work. Negative $\bar{d}$ values express deterioration.
F expresses the statistical significance of the mean differences ($\bar{d}$).
* F at 5 percent of significance is 4.13.
** F at 1 percent of significance is 7.44.

c being the number of combinations of the k illumination levels compared 2 at a time,

$$c=\frac{k!}{2(k-2)!}. \tag{7}$$

The use of the t'-scores has the advantage of presenting the over-all picture of the effect of levels of illumination on different functions in a common frame of reference, taking into account the inter-individual variability of the responses. The means for different levels at illumination are to be interpreted in reference to the level of the grand mean, based on all observations made at 5, 100, and 300 footcandles. The position of a given mean value above or below the grand mean indicates a relative superiority or inferiority of the particular level of illumination intensity. The

TABLE V. Change in several ophthalmological routine tests at three levels of illumination.

	Level of illumination					
	5 ft.-c		100 ft.-c		300 ft.-c	
Function	$\bar{d}$	F	$\bar{d}$	F	$\bar{d}$	F
Abduction power	−0.11	0.34	−0.22	3.24	0.10	0.59
Adduction power	1.35	4.97*	0.88	1.90	0.82	3.10
Vertical divergence	0.01	0.006	−0.04	0.56	−0.01	0.05
Accommodation near point	−0.14	0.47	0.31	2.38	−0.01	0.005
Convergence near point	−0.08	0.27	−0.17	2.24	−0.15	1.11

$\bar{d}$ equals the mean change between 36 pairs of measurements made before and after visual work. Negative $\bar{d}$'s express deterioration.
F expresses the statistical significance of the mean differences ($\bar{d}$).
* F at 5 percent of significance is 4.13.
** F at 1 percent of significance is 7.44.

TABLE VI. Fatigue changes in ophthalmographic functions at three levels of illumination.

	Level of illumination					
	5 ft.-c		100 ft.-c		300 ft.-c	
Function	$\bar{d}$	F	$\bar{d}$	F	$\bar{d}$	F
Movement phase	−0.87	5.94*	−0.15	0.57	−0.18	0.42
Angular velocity	−0.53	2.42	−0.27	0.83	0.03	0.005
No. of movements	−0.18	7.87***	−0.26	30.00***	−0.25	10.71***
No. of movements	−0.19	4.46	−0.24	10.86**	−0.18	1.18
Fixation phase	−1.10	1.32	−1.48	5.48*	−1.36	2.64
Extent of movement during fixation phase	−0.5	2.65	−0.21	0.84	0.38	0.78
Discrepancy between fixation point and target	−0.15	0.18	−0.64	2.40	−0.45	0.93

$\bar{d}$ equals the mean change between 12 pairs of measurements made before and after visual work. Negative $\bar{d}$'s express deterioration.
F expresses the statistical significance of the mean differences ($\bar{d}$).
* F at 5 percent of significance is 4.84.
** F at 1 percent of significance is 9.65.
*** For 36 pairs of measurements F at 1 percent of significance is 7.44.

vertical difference between any two mean t'-scores, read off from the graph, indicates the statistical significance of this difference, expressed in the form of a (t)-test, defined as

$$(t)=\frac{\bar{\Delta}}{(\text{S.D.})_{\bar{\Delta}}}. \tag{8}$$

In the Tables II to VI only the data for 5, 100, and 300 ft.-c are given, each mean being based on 36 observations. In the graphs, the means for 2, 15, and 50 ft.-c, based on 6, 12, and 12 observations, respectively, are added. The actual value of the grand mean as well as the mean scores at different levels of illumination are included in the graphs in order to make the information more complete.

RESULTS

Test Data

Table II shows the results obtained for the performance test, blinking rate, and questionnaire score at 5, 100, and 300 ft.-c. The mean drop in the performance score from the start to the end of the work period was highly significant at all levels of illumination; this means that the visual work was severe enough to result in a real deterioration of performance ("fatigue") within 2 hours. In contrast, the very slight change of the blinking rate during the work task was not statistically significant at any level of illumina-

tion and, consequently, cannot be used as a fatigue index. At all levels of illumination, the difference between the questionnaire score before and after the work was statistically highly significant. All functions listed in Table II show an optimum at 100 ft.-c. There was no glare at the 300-ft.-c level.

Table III shows the significant mean differences ($\bar{\Delta}$) between the three illumination levels (5 *versus* 100, 5 *versus* 300, and 100 *versus* 300 ft.-c); those functions which did not bear a significant relationship to the illumination level are not included. It is of interest that the average blinking rate differentiates well between illumination levels, although the change of the blinking rate from the start to the end of the work cannot be used as a fatigue index.

The mean differences in performance scores and in visual functions at 5 and 100 ft.-c are more marked than the differences between 100 and 300 ft.-c. This is not surprising, since 5 ft.-c are obviously inadequate for the visual task used in these experiments. However, the relative deterioration at 300 ft.-c is of theoretical and practical interest. It can be seen that this deterioration is statistically highly significant for the performance drop, performance range, and the average blinking rate (Table III, last two columns).

The over-all trends in the performance criteria, blinking rate, and questionnaire score as related to different levels of illumination intensity are given in Figs. 1 to 5. Figure 1 shows a pronounced increase of the average performance (output) from 2 to 50 ft.-c; the further slight increase at 100 ft.-c and the decrease at 300 ft.-c was statistically not significant, but it should be noted that actually the performance at 300 ft.-c was slightly below the value at 50 ft.-c. Although for the average performance the optimum at 100 ft.-c was not statistically significant, it fits so well into the general pattern of the other performance criteria (drop and range) that there is little doubt of the reality of a performance optimum above 50 ft.-c but below 300 ft.-c. Both performance drop from start to end of the work period and performance range show a definite optimum at 100 ft.-c (Figs. 2, 3). For the blinking rate average there was a well established optimum at 100 ft.-c (Fig. 4). A less pronounced optimum at the same level was obtained for the questionnaire score

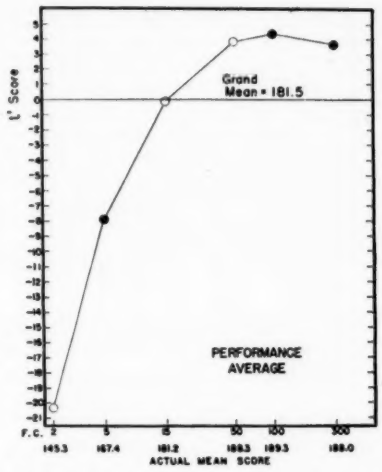


FIG. 1. Effect of the illumination level on the performance average. The values at 5, 100, and 300 ft.-c are based on 36 determinations (solid circles), the values at 2, 15, and 50 ft.-c are based on 6, 12, and 12 determinations (open circles).

(Fig. 5). The distinct deterioration of feeling of well being as reflected in the questionnaire score at 2 ft.-c should be noted; the questionnaire score differentiated significantly between illumination levels only when the intensity became grossly insufficient. On the other hand, the optimum for the questionnaire score at 100 ft.-c, although statistically not significant, should not be considered as a "chance" phenomenon, since it coincides with the pattern of the other functions.

Table IV shows the change in retinal functions resulting from visual work at three levels of illumination. The recognition time for stimuli of threshold size showed progressively increasing "fatigue" changes with decreasing illumination.

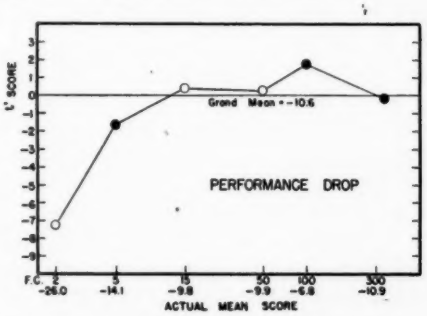


FIG. 2. Effect of the illumination level on the drop of performance during 2 hours of visual work.

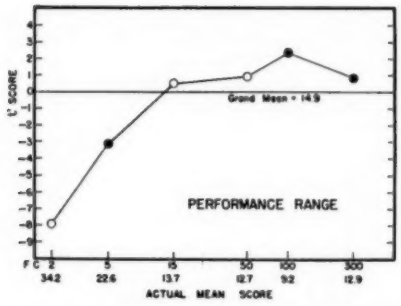


FIG. 3. Effect of the illumination level on the performance range (maximum intra-individual variation of performance during 2 hours of visual work).

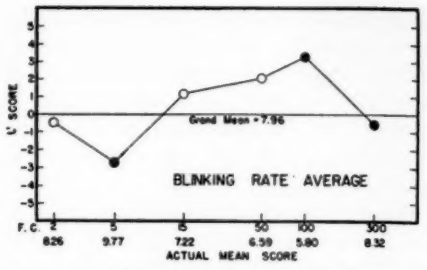


FIG. 4. Effect of the illumination level on the blinking rate during visual work.

However, the mean difference ($\bar{d}$) between values obtained before and after work is statistically significant only at the lowest level (5 ft.-c); the difference ($\bar{\Delta}$) in the degree of deterioration was statistically significant for 5:100 and 5:300 ft.-c (see Table III). The flicker fusion frequency (f.f.f.) showed a statistically highly significant deterioration at all illumination levels, but did not differentiate between the three illumination levels. The brightness discrimination showed no significant changes from start to end of work, and the differences ($\bar{\Delta}$'s) between the three levels of illumination were not significant.

Figure 6 shows the trend of differences ($\bar{d}$) in recognition time except for the value at 2 ft.-c, which might be due to the small number of observations; the trend is continuous up to a level of 300 ft.-c, in contrast to all other functions investigated. Figure 7 shows the trend of changes ($\bar{d}$) in f.f.f.; the optimum (least deterioration) was obtained at 50 ft.-c, and the greatest deterioration at 300 ft.-c; however, the differences between levels of illumination were not statistically significant.

The effects of prolonged, strenuous visual work on the ophthalmological tests are summarized in Table V. Of the functions measured, a significant difference between initial and terminal scores was found only for adduction power, which "improved" significantly at 5 ft.-c. The same tendency was present also at the other levels of illumination, but the changes were statistically not significant. When the data of all three illumination levels were pooled together, the "improvement" was found to be highly significant. The ophthalmological routine tests did not differentiate between illumination levels.

Of the ophthalmographic functions (Table VI), the maximum rate of eye movements (number of movements) showed highly significant fatigue changes at all three levels of illumination for 36 observations and a significant deterioration only at 100 ft.-c for 12 observations. Unfortunately, in all other ophthalmographic functions only 12 measurements were available. Since movement phase, fixation phase, and discrepancy between fixation point and target showed a deterioration at all illumination levels, statistical significance might have been obtained with a larger number of observations. It is of interest that the number of movements does not differentiate between illumination levels, although there were significant fatigue changes at all three illumination levels; none of the other ophthalmographic functions showed significant differences between the three levels of illumination.

Subjective Preferences

For the realistic evaluation of the optimum level of illumination the subjective preferences are also important. The questionnaire score

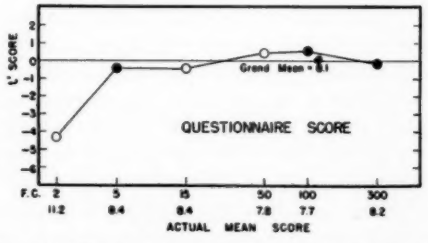


FIG. 5. Effect of the illumination level on subjective discomfort, scored on the basis of 10 items of a questionnaire.

(Table II; Fig. 5) was based on the negative effects (discomfort) produced by the visual task. In contrast, a preference score expresses the positive selection of the most agreeable illumination level. The preference score was determined, after all of the experimental work was completed, for the six experimental subjects as well as for 14 untrained subjects. In five of the individual booths 5 different illumination levels were arranged (5, 15, 50, 100, and 300 ft.-c) in random order. The subjects did not know the actual levels of illumination used.

They looked at the moving letters in the same way as during the work tests and ranked the booths in the order of preference from "1" (most preferred) to "5." The subjects were tested one at a time. The length of time the subjects wanted to spend to make their ratings was not limited.

We are aware of the limitation of this procedure. It is known[19a] that the pre-adaptation level is important for the preference rating. In passing from low levels of illumination to high levels, the preference rating might be shifted toward lower illumination; the reverse sequence might reverse the preferences. Thus it may appear that an uncontrolled source of error might have modified the ratings.

However, we believe that this factor was negligible. The extreme illumination levels were eliminated quickly and with ease, and the greatest part of the time was spent with the rating of two or three medium illumination levels. The difference between these levels was comparatively small, reducing the importance of this error. Also, the difference between room illumination and booth illumination played probably only a minor role as interference factor. For the testing situation, the room illumination was irrelevant since the booths were large enough to represent their own surrounding illumination. The subjects went rapidly from one to another booth, so that there was no time to adapt to the room illumination. Furthermore, it should be noted that in the averages the different sequences at which the illumination levels were inspected by the different subjects must have cancelled out to a large degree any preadaptation trends.

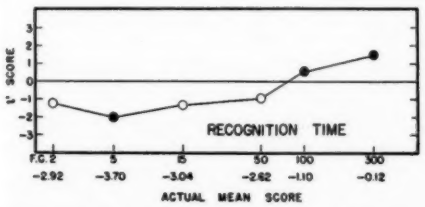


FIG. 6. Effect of the illumination level on the change of recognition time for threshold size stimuli, tested before and after the visual work. The grand mean was −1.64 sec.

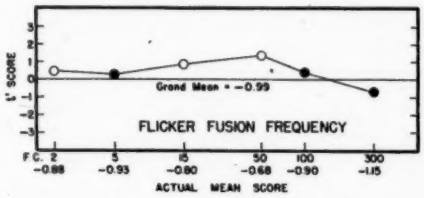


FIG. 7. Effect of the illumination level on the change in the fusion frequency of flicker, tested before and after the visual work.

The estimates were made separately for the three types of lamps. The results are given in Table VII. It can be seen that the 100 ft.-c is the preferred level, and that this optimum is statistically highly significant. This is true for all lamps and for both trained and untrained observers. It is interesting that the subjective preference level coincides with the apparent physiological optimum. This does not mean, of course, that this will always be the case; certainly, it would be hazardous to base illumination level recommendations on subjective preferences only.

DISCUSSION

Of the 23 items which were investigated and which served as criteria of performance and of the visual functions, only comparatively few show significant fatigue changes as a result of the very strenuous visual task. The changes in the performance criteria (average performance, performance drop, and range) are excellent indices of visual strain. The f.f.f., recognition time, maximum rate of eye movements, and subjective complaint score proved to be sensitive to the stress of visual work. It is surprising that the ophthalmological tests which are used routinely for investigation of patients who complain of eye strain failed to show any fatigue trend; in fact,

[19a] M. A. Tinker, "Effect of visual adaptation upon intensity of light preferred for reading," Am. J. Psychol. **54**, 559 (1941).

TABLE VII. Average preference scores for 5 levels of illumination intensity. The smaller the score, the more preferred the condition.

Subjects	Lamp	Illumination level ft.-c					significance		
		5	15	50	100	300	50:100	100:300	50:300
6 trained	Frosted	5	3.7	2.2	1.7	2.5	**	**	no
14 untrained		5	3.9	2.4	1.3	2.5			
6 trained	Natural white	5	3.7	2.2	1.8	2.3	**	**	no
14 untrained		5	3.7	2.5	1.3	2.5			
6 trained	Verd-A-Ray	4.8	3.5	2.2	1.8	2.8	*	**	no
14 untrained		4.9	3.9	2.3	1.5	2.5			

* Indicates significance of the mean differences at the 5 percent level.
** Indicates significance at 1 percent level. In testing the significance of the differences the combined sample of subjects was used ($N=20$).

the adduction power shows a rather consistent improvement which reaches statistical significance at the 5-ft.-c level. The improvement of a visual function during hard visual work in unexpected and deserves comment. It is known that muscle coordination and mechanical efficiency improve up to a certain point in the course of some types of muscular work (literature summarized by Simonson.[20] The length of this adaptation period varies in different types of work and has not been studied for eye muscle functions, but it seems possible that the adduction power of the eye muscles was, at the end of the 2-hour work period, still in the phase of improvement (adaptation), while other visual functions, including eye muscle coordination as recorded by the ophthalmograph, showed a definite deterioration. We may assume that the functions which exhibited definite fatigue trends during the work period were more directly affected by the visual strain; it would depend on the duration of the work task whether deterioration would result also in those functions less intimately related to the type of visual strain present in the given work test. We tested and confirmed this hypothesis by extending the work period to 4 hours; under these conditions, a deterioration occurred also in the adduction power.

Adaptation, i.e., improvement trends during visual work, have been observed before, for instance, acuity and brightness discrimination improved during visual work involved in sorting of plates (Ströer).[16] That the changes of the various visual functions show such great differences in time and magnitude and even sometime in direction is of great importance and demonstrates the impossibility to interpret or define "visual fatigue" in terms of only one variable. Discrepancies in the different criteria of an adequate illumination intensity have been described by other investigators.

Freeman,[15] using visibility, speed of reading, and accuracy in proofreading as criteria, found that continuous reading for one-and-half hours under 0.5-ft.-c illumination resulted in fatigue trends; at 1 ft.-c no measured evidence of fatigue was obtained and this level would be, by definition, adequate. On the other hand, the legibility curves of the unfatigued eye reached a plateau, in the majority of adult subjects, at 5 ft.-c. According to the legibility data, a level of 5 ft.-c or more would have to be recommended. Visibility determinations cannot be used for prediction of fatigue trends, or *vice versa*. Ströer[16] found discrepancies between visual acuity and performance, and between number of errors and output.

The failure of abduction, adduction, and convergence to show fatigue trends contrasts with the findings of Luckiesh and Moss.[21] These authors reported for a much less strenuous and less prolonged task (reading for 1 hour) not only a consistent deterioration during the task, but even a differentiation between two illumination levels (1 ft.-c and 100 ft.-c). Their type of work differs from ours, but there is no question that our task was much harder than ordinary reading.

Comparison of Tables V and VI shows that the ophthalmographic approach holds more promise to detect fatigue changes than the ophthalmological tests. It is possible, of course, that these might be more sensitive in patients with weak eye muscles or disturbance of eye muscle coordination.

It may be noted that of the three retinal functions tested, two (recognition time and f.f.f.) show fatigue trends while brightness discrimination does not. The reason might be that brightness discrimination was tested 15 minutes after the end of work, while recognition time and f.f.f. were tested immediately after the subjects finished the two hours of work. The visual acuity

[20] E. Simonson, "L'adaptation au travail physique," Le Travail Humain **4**, 129 (1936).

[21] M. Luckiesh and F. M. Moss, "Fatigue of convergence induced by reading as a function of illumination intensity," Am. J. Ophthal. **18**, 319 (1935).

threshold itself did not change significantly during the visual task. This might be due in part to the limits of our testing procedure. The steps in the size of the various dots might have been too crude to detect minor changes of visual acuity, although they were much finer than the steps in letter size in the usual visual charts.

The increased intra-individual variability of performance resulting from visual strain and measured as the performance range for an individual is both of theoretical and practical interest. The results corroborate findings reported by Ströer[16] and by Schulz.[22] The increase of intra-individual variability may be associated with the increase of inter-individual variability as shown elsewhere.[13] Both the increase of inter- and intra-individual variability would interfere with the maintenance of a smooth performance, especially in timed and coordinated team work, characteristic of much of the modern industrial operations.

We regard the demonstration of an optimum at 100 ft.-c in most of the visual functions which show significant fatigue trends as our most important result. The level of the optimum is below the minimum recommendations of the I.E.S., which indicates a level above 100 ft.-c for strenuous visual work and is definitely below the implied "desirable" level. The existence of an optimum which was to be expected from physiological considerations is not compatible with the I.E.S. general conception of illumination requirements. An excess of illumination, beyond a certain level, may actually produce a deterioration of performance and deteriorative changes in visual functions. Observations of an optimum illumination level for various visual functions have also been made by Ströer,[16] by Ferree and Rand,[23] and by Weston.[14] Future recommendations should be given in terms of an optimum rather than a "minimum" level. According to our results, there is reason to believe that the optimum will be substantially below the levels which are assumed to be desirable at the present time. It is true that many of the functions tested did not show any optimum nor any significant trend with illumination level or fatigue. Such inert functions obviously cannot be used for discussion of illumination effects. The only exceptional function exhibiting a significant trend with the illumination level, not consistent with an optimum, is the recognition time of threshold stimuli. This function shows a continuous improvement trend from 2 to 300 ft.-c.

The relationship of this function to illumination intensity is obviously quite exceptional. It is unfortunate that the present recommendations are to a great extent based on visual acuity determinations. This explains, in part, the high recommendations of the I.E.S. and Luckiesh and Moss.[12] In fact, if we had limited our work just to the recognition time for threshold stimuli (Fig. 6) we would have arrived at the same conclusion. Yet it would be obviously an incorrect procedure, as the "recommended" level would be based on the behavior of an exceptional function. Criticism of using visibility and visual acuity as a basis for illumination level recommendations has been offered by Tinker[24] and Ryan.[25]

In general, illumination engineers tend to exaggerate the effect of variations in the illumination level. This is shown in the present study (1) by the location of the optimum in a severe visual work at a relatively low level, and (2) by the magnitude of variations in the illumination level necessary to produce significant differences. The trend of changes in reference to the illumination level in several functions (f.f.f., maximum rate of eye movements, questionnaire score between 5 and 300 ft.-c) is very slight as compared to the highly significant changes taking place during the course of work at all illumination levels. Such observations have been repeatedly made in the past, for instance by Ströer[16] in regard to changes of visual acuity and brightness discrimination, by White, *et al.*[26] for the decrement of visual acuity of letter separators tested in the late afternoon as compared with the values obtained in the morning, and by Ferree and Rand[23] for the ability

[22] H. Schulz, "Die Ermüdung des Auges," Industr. Psychotechnik **2**, 5 (1925).

[23] C. E. Ferree and G. Rand, "The efficiency of the eye under different conditions of lighting: the effect of varying the distribution factors and intensity," Trans. Illum. Eng. Soc. **10**, 407 (1915).

[24] M. A. Tinker, "Illumination standards," Am. J. Pub. Health **36**, 963 (1946).

[25] T. A. Ryan, "Work and effort: the psychology of efficiency," (in press).

[26] L. R. White, R. H. Britten, J. H. Ives, and L. R. Thompson, "Studies in illumination II. Relationship of illumination to ocular efficiency and ocular fatigue among the letter separators in the Chicago post office," Public Health Bull. No. 181, Washington (1929).

to maintain fixation without blurring. Also, Hoffman[27] in his analysis and criticism of Luckiesh and Moss' findings comes to similar conclusions. The only possible interpretation is that fatigue trends resulting from visual work must be produced by components which are common at all illumination levels. The illumination level is by no means the all-important variable for the development of visual fatigue.

Naturally, the question arises as to how the discrepancy between our results and the high illumination recommendations of the I.E.S. can be explained. The I.E.S. reports concerning the various industries are based mainly on practical experience. Although such information is valuable and should by no means be discarded, scientific approach demands a more rigorous proof for the claims which have been made. It is amazing in how many studies on visual performance the established rules and precautions for the scientific study of human performance are neglected. In most investigations the output depended on other than visual functions, such as manual skill or verbal intelligence. Also, the statistical treatment of the results in many papers leaves much to be desired. The results were frequently marred by prolonged practice or training trends or a change in motivation. The subjects were made illumination conscious by the experiments. As a rule they knew that the illumination level was changed, and no control experiments with simulated changes, such as exchange of lamps of equal wattage, were made. This experimental situation is comparable to experiments on the effects of drugs or vitamins in human beings without placebo controls, an obviously unacceptable procedure.

In Great Britain the opinion is prevalent that industrial work involving fine visual discrimination and perception of details requires from 10 to 25 footcandles. For work involving assembly of large parts 6 to 10 footcandles are considered adequate, whereas for high precision and prolonged work the recommended level would go up to 50 footcandles, and possibly higher. These figures which are markedly lower than those of the I.E.S. code were presented by Oakley[28] and are based on the codes of industrial lighting issued by the (British) Illuminating Engineering Society and the Electrical Lighting Manufacturers Association, London. Weston and Taylor[29] showed that the rate of output in typesetting by hand and the ratio of total errors to output reached an efficiency comparable to that obtained under good daylight illumination, at about 20 ft.-c.

We have been impressed with the fact that much of the older work, performed 20 to 30 years ago, has been ignored in the recent American literature, although the investigations were, in regard to experimental design and thoroughness, as good if not better than some of the more recent studies. Only a few examples can be given. Thus the National Education Association set up a Committee for the Illumination Requirements of School Buildings, the report of which was published by F. N. Freeman;[15] the Committee arrived at the conclusion that the standards of the I.E.S. at that time were higher than necessary, which is interesting, since the standards 25 years ago were much lower than those presently recommended. One of the most extensive studies on the effect of illumination intensity was sponsored by the National Research Council in the early twenties. Snow's report[30] shows clearly that the rise in output, observed in some types of work, was not a simple function of illumination; in other types of work, the variations of output showed no relationship whatever to the level of illumination.

Working with Landolt broken rings, Weston[14] found the optimum of performance at 20 foot candles with a size of the gap corresponding to a visual angle of 10 minutes, the same as in our test. The lower optimum level in Weston's test is probably due to two factors: (a) the movement of the letters in our work test which increases the severity of the task and (b) the configuration of the Landolt rings which is markedly simpler than that of the letters. Weston did not arrive at the concept of a physiologically optimal level of illumination intensity, and simply stated that

[27] A. C. Hoffman, "Luckiesh and Moss on reading illumination," J. App. Psychol. **31**, 44 (1947).

[28] C. A. Oakley, *Men at Work* (University of London Press, London, 1946), p. 301.

[29] H. C. Weston and A. K. Taylor, "The relation between illumination and efficiency in fine work (typesetting by hand)," Joint Rep. Industr. Fatigue Res. Board and Illum. Res. Comm., (London) (1926).

[30] C. E. Snow, "A discussion of the relation of illumination intensity to productive efficiency," Tech. Eng. News **8**, 257 (1927).

with gaps larger than 6′ in size the performance at illumination levels above 2 ft.-c becomes practically independent of illumination.

Luckiesh and Moss[12] used changes of blinking rate, nervous muscular tension, and heart rate to support their claim for high illumination levels. However, subsequent investigators have largely failed to confirm their results (Tinker;[31] McFarland *et al.*;[32,33] Bitterman[34]). It is curious that Luckiesh and Moss interpret a decrease of heart rate as an index of strain, contrary to all available physiological evidence; physiological strain tends to increase, not decrease, the heart rate.

We wish to emphasize that the specific conclusions arrived at in this study are limited to our type of visual task and subjects. Obviously, further work along these lines is necessary for other types of visual work. The question arises how far our results may be generalized. We have naturally standardized a majority of factors and varied only the illumination level. In general, our visual work exceeds in severity most tasks found under actual working conditions. Therefore, the development of fatigue would occur more slowly in most types of industrial inspection work concerned with discrimination of fine details. It seems reasonable to expect that changes observed in our 2-hour work task may be comparable to fatigue developed within 6 to 8 hours of fine visual work in industry.

It is apparent that in determining the optimal level of illumination, the specificity of the particular job has to be kept in mind, including such factors as dimensions of the critical details (visual angle), intensity of illumination at the level of the object, contrast between the object and the background, reflection factors, speed and steadiness of the movement of the object, duration of exposure, duration of the visual work, characteristics of the details to be recognized (definite contour), color of the object, and state of adaptation of the eye.

[31] M. A. Tinker, "Validity of frequency of blinking as a criterion of readability," J. Exp. Psychol. **36**, 453 (1937).

[32] R. A. McFarland, A. H. Holway, and L. M. Hurvich, "Studies of visual fatigue," Harvard Univ. Graduate School of Business Administration (1942).

[33] R. A. McFarland, C. A. Knehr, and C. Berens, "Metabolism and pulse rate as related to reading under high and low levels of illumination," J. Exp. Psychol. **25**, 65 (1939).

[34] M. E. Bitterman, "Heart rate and frequency of blinking as indices of visual efficiency," J. Exp. Psychol. **35**, 279 (1945).

ACKNOWLEDGMENT

We wish to thank Dr. Ancel Keys, Director of the Laboratory of Physiological Hygiene, for the help, encouragement, and criticism given us at all stages of the experiment, and Dr. Erling Hansen, head of the Division of Ophthalmology, University of Minnesota, for his interest and advice. The study was made possible by the full cooperation of the men who served as subjects: Hughes Browne, Reed F. Landis, Peter S. Olmsted, Carl E. Patterson, William Steinberger, and Robert Villwock.

We wish also to thank Mr. W. P. Thompson for his careful statistical work and Mr. O. S. Levi, Director of Research, Verd-A-Ray Corporation, Toledo, Ohio, for calibration of the lumen output of all lamps used in this study.

Photographic Granularity and Graininess. IV. Visual Acuity Thresholds; Dynamic *versus* Static Assumptions*

LOYD A. JONES AND GEORGE C. HIGGINS
Kodak Research Laboratories, Eastman Kodak Company, Rochester, New York

(Received October 30, 1947)

The distances at which a series of rectangular test objects are visible were determined experimentally. These test objects were black opaque rectangles having variable length-width ratios and were viewed against a transluminated background. Using these data, and an assumed nystagmic motion of the eyeball, the *summated temporal illuminance gradients* on the retina were computed. These values were found to be essentially constant, even though the length-width ratios of the test objects varied from 1 to 4050. Assuming a motionless eyeball, that is, perfect fixation, the summated values of ΔI were computed for the same observational conditions, ΔI being defined as the difference between the average illuminance incident on adjacent cones in the retinal mosaic. The values of summated ΔI are not constant. Since all of these test objects were *just visible* at the distance measured, they may be considered as being equal to each other with respect to their ability to excite a threshold neural response. It seems reasonable to conclude, therefore, that since the values of the summated temporal gradients are constant they are the most significant indices of the relative magnitudes of the neural responses. In our opinion these results lend considerable support to the assumption that the perception of inhomogeneities of luminance in the visual field is dependent directly upon the temporal illuminance gradients to which the retinal receptors are subjected by virtue of the movement of the image with respect to the retinal mosaic.

IN the third paper[1] of our series on granularity and graininess we presented data on the visibility of rectangular test objects having a varying length-width ratio. The distribution of illuminance on the retina, when the square test object is at the threshold of visibility, was shown in Fig. 10 of this paper. Curves showing the amplitudes of illuminance to which retinal cones are subjected during one-half cycle of an assumed nystagmic motion across the illuminance gradient were shown in Fig. 11. Curves showing the temporal gradients on the retinal cones, as computed from the slopes of the curves in Fig. 11, were shown in Fig. 12.

It seemed that it might be instructive to calculate the summated average temporal gradients for all the cones taking part in the perception of the square test object and to compare this sum with the corresponding values calculated for the other test objects at their thresholds of visibility. We are aware that there is little similarity between the long, narrow line test objects, and the elements of a granular structure in a developed photographic material. However, an understanding of the manner in which the visual system functions in the perception of these test objects is of importance in the development of a general theory of the visual process and may assist us in finding a method of measuring granularity which will yield values in agreement with graininess measurements.

METHOD OF CALCULATION AND RESULTS

The distribution of illuminance along a line on the retina which is perpendicular to the long dimension of the image of the test object was plotted for each test object when it is at the threshold of visibility. These curves were then corrected to account for the effect of scanning the diffraction image with a cone which has a finite size. In all cases the difference between the actual diffraction image and the distribution of illuminance, as measured with a scanning area corresponding to the size of the central retinal cones, was very small. The curves are all similar to the one shown in Fig. 1, which represents the distribution of illuminance in the image of the test object having a length-width ratio of 100. The retinal cones are represented in this figure by the cross-hatched areas.

The amplitudes of the illuminance to which the cones were subjected during their nystagmic motion across the spatial illuminance gradient were computed from the curves for the distribu-

* Communication No. 1164 from the Kodak Research Laboratories.

[1] L. A. Jones and G. C. Higgins, "Photographic granularity and graininess. III. Some characteristics of the visual system of importance in the evaluation of graininess and granularity," J. Opt. Soc. Am. **37**, 217 (1947).

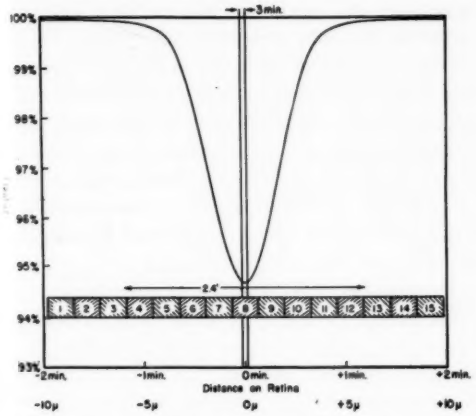


FIG. 1. The distribution of illuminance on the retina, at the limit of visibility, for a rectangular test object having a length-width ratio of 100.

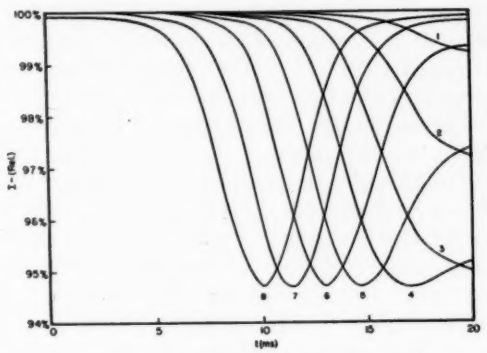


FIG. 2. Curves showing the amplitudes of illuminance to which retinal cones are subjected during one-half cycle of the physiological nystagmus when the rectangular test object having a length-width ratio of 100 is at the threshold of visibility.

tion of illuminance. It was assumed that the cones moved with a simple harmonic motion having a frequency of 25 cycles per second and an amplitude of 2.4 minutes, as reported by Adler and Fliegelman.[2] These values of illuminance were plotted as a function of time. The curves obtained for the test object having a length-width ratio of 100 are shown in Fig. 2; the numbers on the curves refer to the numbers on the cones in Fig. 1. The ordinates give the values of average illuminance on each cone relative to the illuminance that would be on the cone if the dark test object were removed, and the abscissa represents the time in milliseconds reckoned from zero at the instant when the cones are at the extreme left position in their vibratory cycle.

It should be noted that the curves in Fig. 2 apply to only one-half cycle of the motion and that the curves have been drawn for only one half of the cones acting. Because of symmetry conditions, the curve for Cone 7 is the same as for Cone 9. Likewise, the curves for Cones 6 and 10, 5 and 11, etc., are identical. In a like manner, the curves for the second half of the cycle are the same as those for the first half.

The gradients of the curves showing the variation of illuminance on the cones as a function of time (see Fig. 2) are a measure of the *rate* at which the change in illuminance occurs. These temporal gradients, dI/dT, were measured and plotted as a function of time. The curves obtained from Fig. 2 for the test object having a length-width ratio of 100 are shown in Fig. 3. Each curve in this figure represents the temporal illuminance gradients experienced by a single cone. There are at least fifteen cones which are receiving a temporal gradient during a complete cycle of the nystagmic motion.

While the curves in Fig. 3 show the temporal gradients only on Cones 1–8, as pointed out previously, the curves for Cones 9–15 (see Fig. 1) are the same as for Cones 7 to 1. A set of curves similar to those shown in Figs. 1 to 3 were computed for each test object. These curves are the basis for the calculation of the summated average temporal gradients on the retinal cones at the limit of visibility for the various test objects. For a more detailed description of the method of obtaining these curves, the reader is referred to our previous paper.[1]

The data for the computations of the summated average gradients are given in Table I. The values of the length-width ratio (L/w) are given in the first column. The angle (α) in minutes subtended by the long dimension of each test object at the threshold of visibility is given in the second column and the angle (β) in seconds subtended by the short dimension is given in the third column. These values of α and β were taken from the smooth curve obtained

[2] F. H. Adler and F. Fliegelman, "Influence of fixation on the visual acuity," Arch. f. Ophthalm. 12, 475 (1934).

by plotting the experimentally determined values of α as a function of the experimentally determined values of β. The observed values are given in Table III of our previous paper.[1]

The values for the average center distance $(CD)_{Av}$ between adjacent cones taking part in the process of seeing each of the test objects are shown in the fourth column of Table I. These were taken from the curve in Fig. 2 of our previous paper[1] which represented the variation in the center distance between cones as a function of the angular distance from the center of the fovea. The angular distance from the center of the fovea used in obtaining these average *CD*'s was taken from the angular length (α) of the long dimension of the test object plus the extension of this length by reason of diffraction and the nystagmic motion of the image across the cones. For a test object having a small length-width ratio, the geometric image covers only a very small part of the cones experiencing a temporal illuminance gradient. The effect of diffraction and nystagmic motion on the number of cones receiving an illuminance gradient can be seen quite clearly from Fig. 4. In this figure the geometric image is indicated by the small dark square in the center. The concentric circles surrounding this square represent lines of equal illuminance, the values of which are given in percent of the maximum illuminance. The large dotted circle (*A*) represents the limit of the retinal area which will move into the area of illuminance gradient by virtue of the nystagmic motion of the retina as a whole across the image on the cone mosaic. The value of $(CD)_{Av}$ for the square is computed on the basis that all the cones in the area surrounded by the dotted line are acted upon by a temporal illuminance gradient.

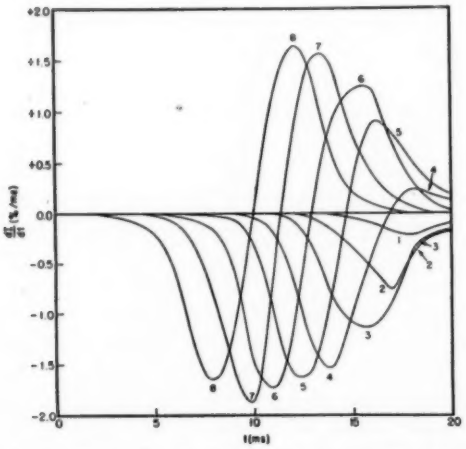


FIG. 3. Curves showing the rate of change of illuminance incident on the retinal cones during one-half cycle of the physiological nystagmus when the rectangular test object having a length-width ratio of 100 is at the threshold of visibility.

The number of cones (N) covered by the *geometric* image of the long dimension of the test object is shown in column 5 (Table I). These values were obtained as the quotient of the angular length of the geometric image (α) and the angular width of the $(CD)_{Av}$ between cones. ($(CD)_{Av}$ in column 4 is in μ and 1μ is equivalent to an angular subtense of 0.2 minute.)

As the cones move back and forth across the illuminance gradient along a line perpendicular to the long dimension of the test object, as shown in Figs. 1 to 3, several cones are acted upon by the temporal gradient. The number of cones involved depends upon the size of the cones, the amplitude of the vibratory motion of the cones relative to the image, and the characteristics of the illuminance gradient. Because of the nature of the trailing off at the ends of the illuminance gradient (see Fig. 1), the number of cones considered to be taking part in the visual process will depend on the limit set as the lowest value of the temporal illuminance gradient which will be considered as a liminal stimulus. Since it is known that the line having a length-width ratio of 4050 can be seen, we can assume that the lower limit for the liminal temporal gradient must be at least as low as the maximum gradient experienced at the limit of visibility of this test object. The maximum temporal gradient on any cone during the perception of the test object which is 4050 times as long as it is wide, is 0.26 percent per millisecond. Fortunately, in the calculation of the summated average temporal gradients, the choice of a limiting gradient is not a serious problem, since it is clear that those gradients which are so small that they make no significant contribution to the total gradient can be neglected. Employing the criterion that the gradient must be large enough to make a significant contribution to the total, the number of

cones (n) acting in a single row perpendicular to the long dimension of the test object is that shown in column 6 of Table I. An examination of the gradients on those cones which were not included in the computation of the value of n indicated that the maximum for any cone which had been omitted from the calculations was 0.016 percent per millisecond. That is, all cones which were omitted received a maximum temporal gradient which was less than one tenth of the maximum gradient for the longest line. In addition, it was found that the maximum temporal gradient acting on the limiting cones was 0.24 percent per millisecond. It is interesting to note that the arbitrary method employed in determining the limiting minimum temporal gradient gave values which were in good agreement with the theoretical limit for a liminal stimulus.

The curves showing dI/dt as a function of t, such as Fig. 3, represent the temporal gradients experienced by a single row of cones perpendicular to the long dimension of the test object during one-half cycle of the nystagmic motion or 20 milliseconds. The area under each curve was measured and the sum of the areas for all the cones divided by 20 is taken as the summated average temporal gradient for this row of cones perpendicular to the long dimension of the test object. These values (Z) are given in the seventh column of Table I. The values of Z have the dimension of dI/dt, where I is the illuminance relative to the maximum, and t is in milliseconds.

In the case of a square test object, the illuminance pattern at the limit of visibility is essentially circular (see Fig. 4), and the summated average temporal gradient is independent of the direction of the motion of the cones. On the other hand, for those test objects which have one dimension longer than the other, the value of the temporal gradient depends upon the direction of motion of the cones with respect to the long dimension of the pattern. For the rows of cones at some distance from the ends of the pattern, the random motion can be resolved into components perpendicular and parallel to the long side. If the cones move in the direction represented by the line A in Fig. 5, and if this direction makes an angle, α, with the line B which is perpendicular to the long side of the

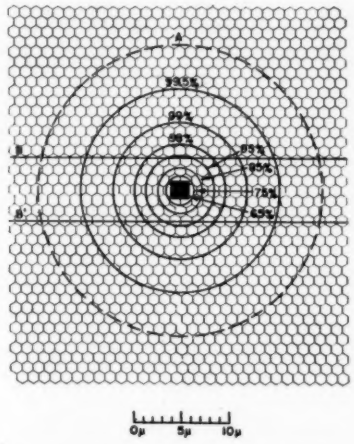


FIG. 4. Diagram showing the distribution of illuminance on the retinal mosaic at the limit of visibility for the square test object.

image of the rectangular test object, then the component of velocity in this perpendicular direction is the cosine α. The average value of this component of the motion perpendicular to the long side is given by the integral of cosine α, and is numerically equal to 0.64. That is, each row is only 64 percent as efficient in producing temporal gradients when moving across the central portion of the illuminance pattern for a long narrow line as it is when moving across the essentially circular pattern for the square.

From calculations based upon Fig. 4, it was found that for this circular illuminance pattern the total summated average temporal gradient for all rows was 1.9 times the summated average temporal gradient for the central row. These computations were made by calculating the summated average temporal gradient for each row across the pattern shown in Fig. 4 in the

TABLE I. Summated average temporal gradients at the limit of visibility of rectangular test objects varying with respect to the length-width ratio.

L/W	$\alpha(')$	$\beta('')$	$(CD)_{kl}$ (μ)	N	n	Z (%/ms)	$1.9Z$	$0.64NZ$	ΣG_{kl} (%/ms)
1	0.36	20.40	1.12	1.6	17	32.2	61	—	61
5	.81	9.36	1.15	3.5	17	15.6	29.6	34.9	65
10	1.33	7.01	1.17	5.6	16	11.8	22.4	42.2	65
20	2.00	5.28	1.21	8.1	16	9.8	18.6	50.8	69
50	3.57	4.25	1.25	14.3	16	6.2	11.8	56.7	69
100	5.21	3.05	1.31	20.0	15	4.6	8.7	58.9	68
218	7.45	2.06	1.33	28.0	15	3.2	6.1	57.3	63
1082	18.8	.83	1.49	63.3	13	1.5	2.9	60.8	64
2610	31.9	.74	1.59	99.9	12	.95	1.8	60.7	63
3410	39.3	.70	1.70	115.3	11	.80	1.5	59.0	61
4050	45.7	.68	1.80	126.2	11	.75	1.5	60.6	62

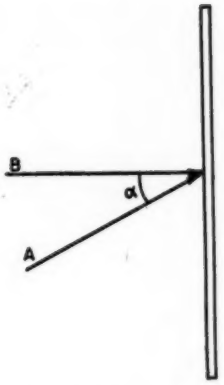


FIG. 5. Schematic diagram for the resolution of the random motion of the eye into the component perpendicular to the long dimension of the test object.

same manner as the summated average gradient was computed for the central row. This involved computing the three curves similar to Figs. 1 to 3 for each row of cones. The value for the total summated average temporal gradient for the square test object is given in the last column of Table I.

The illuminance patterns on the fovea, for those test objects having a length-width ratio greater than unity, are similar to that shown in Fig. 6 for the test object having a length-width ratio of 100. The geometric image is indicated by the dark rectangle in the center of the figure. The continuous lines surrounding this rectangle represent the loci of the points receiving equal illuminance, and the broken line represents the limit of the area containing cones receiving an illuminance gradient by virtue of nystagmic motion. If section A, which is equal in length to the geometric image, is removed, the two ends, when placed together, represent a symmetrical pattern much like that for the square test object (Fig. 4). The contribution of these ends to the total summated average temporal gradient will therefore be approximately $1.9Z$, where, as stated previously, Z is the summated gradient for a single row of cones perpendicular to the long dimension of the test object. Each row of n cones included in the section A of Fig. 6 makes a contribution of $0.64Z$ to the total summated average gradient, since, as pointed out, these rows are only 64 percent as efficient in producing temporal gradients as those moving across the circular part of the illuminance pattern. The total contribution of the cones included in section A is, therefore, $0.64NZ$, since there are N rows of cones in this distance.

The total summated average gradient is the sum of the component produced by section A and that produced by the symmetrical end sections. The value of the total summated average temporal gradient, $\sum G_{Av}$, is

$$\sum G_{Av} = 1.9Z + 0.64NZ. \qquad (1)$$

The values of $1.9Z$ are given in the eighth column of Table I and the values of $0.64NZ$ are in the next column. The values of the summated average temporal gradient, $\sum G_{Av}$, are in the last column of this table.

The values for the summated average temporal gradients, $\sum G_{Av}$, as given in Table I, all apply to the conditions existing at the threshold of visibility for the various test objects. It is interesting to note that these values, which may be considered measures of threshold stimuli, are essentially constant within the limits of experimental error. Since the number of cones being acted upon by a temporal illuminance gradient varies with the nature of the test object, it is interesting to examine this number for each test object when it is at the limit of visibility. At the limit of visibility for the square test object, all the cones included in the area surrounded by the dotted circle in Fig. 4 would be acted upon by a temporal illuminance gradient if the interval of time was sufficiently large to include random motion in all directions. On the other hand, during one-half cycle of nystagmic motion, no cones outside of the area bounded by the lines B and B' and the dotted circle would be acted upon by a temporal gradient equal to, or greater than, that for the cones included in the computations of the summated average temporal gradients. Even in this restricted area many of the cones near the ends of the rows adjacent to the lines B and B' will not be acted upon by temporal gradients as large as those taken as the lower limit for the temporal gradient. This analysis was repeated for all test objects and the total number of cones considered to be taking part in the visual process at the threshold of visibility for each test object was computed.

These values are given in the second column of Table II.

The luminance of the field employed when making the judgments of the angles subtended by the rectangular test objects at the threshold of their visibility was 26 ft.-L. The data reported by Reeves[3] indicate that at a field luminance of 26 ft.-L the iris of the eye should be approximately 3 mm in diameter. Working at this aperture, the luminance of 26 ft.-L should produce an illuminance on the fovea of about 0.52 ft.-c. The values of $\sum G_{hv}$ in percent per millisecond can therefore be converted into approximate values of foot-candles per millisecond. The values of $\sum G_{hv}$ in terms of foot-candles per millisecond are given in column 3 of Table II. The quotient of $\sum G_{hv}$ and the total number of cones, Y, taking part in the perception of the test object at the threshold of its visibility are given in the last column of Table II. These values represent the average temporal gradient per cone, since both $\sum G_{hv}$ and Y were computed for one-half cycle of the nystagmic motion. As pointed out previously, the values of the summated average temporal gradient, and consequently the average temporal gradient per cone, are the same for one-half cycle of the motion as for any other multiple of this interval. The values of the average gradient per cone in terms of foot-candles per millisecond appear to be quite small, but this is entirely the result of the very small unit of time employed in computing the temporal gradients. Temporal gradient values in terms of foot-candles per second, for example, might furnish more convenient units for measuring $\sum G_{hv}$.

Inasmuch as the summated average temporal gradients are essentially constant when measured at the threshold of visibility of the rectangular test objects having widely different length-width ratios, the question arises as to the results which would be obtained by assuming static conditions. Employing the symmetrical properties of the illuminance patterns, such as those shown in Figs. 4 and 6, the values of total summated ΔI have been computed for all test objects. The values of total summated ΔI, $T\Sigma\Delta I$, in foot-candles are given in the last column of Table III.

[3] P. Reeves, "The response of the average pupil to various intensities of light," J. Opt. Soc. Am. 4, 35 (1920).

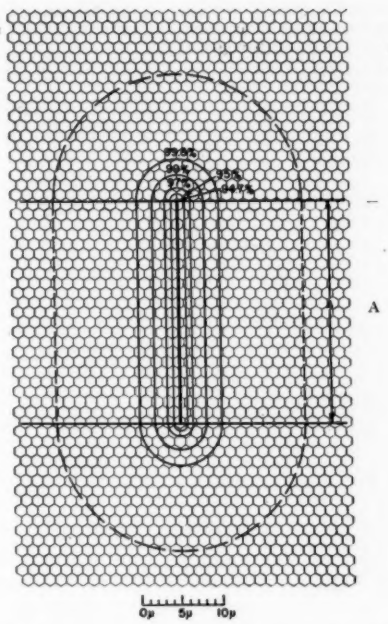


FIG. 6. Diagram showing the distribution of illuminance on the retinal mosaic at the limit of visibility for the rectangular test object having a length-width ratio of 100.

The illuminance patterns were considered as being made up of several contiguous annuli one cone wide, each annulus having the same form as the solid lines in Figs. 4 and 6. Since these solid lines represent points of uniform illuminance, all cones in each annulus receive the same illuminance. Therefore, the product of ΔI, the illuminance difference between one annulus of cones and its neighboring annulus, and the number of cones in the annulus, represents the summated ΔI for this ring of cones. The value of the total summated ΔI, $T\Sigma\Delta I$, is simply the sum of the summated values for each annulus in the pattern.

The number of cones, y, considered to be taking part in the process of visual perception at the limit of visibility of the various test objects on the basis of a static condition of the eye are given in the second column of Table III. The lower limit employed for the value of ΔI in these calculations was about 0.001 ft.-c. While a lower limiting value for ΔI would increase the value for the number of cones acting, it would not change the value for the total summated

TABLE II. Average temporal gradient per cone at the limit of visibility of rectangular test objects varying with respect to the length-width ratio.

L/W	Y	ΣG_{kt} (ft.-c/ms)	$\Sigma G_{kt}/Y$ (ft.-c/ms/cone)
1	87	0.32	0.0037
5	147	.34	.0023
10	172	.34	.0020
20	212	.36	.0017
50	311	.36	.0012
100	377	.35	.0009
218	497	.33	.0007
1082	889	.33	.0004
2610	1260	.33	.0003
3410	1324	.32	.0002
4050	1444	.32	.0002

TABLE III. Total summated ΔI at the limit of visibility of rectangular test objects varying with respect to the length-width ratio.

L/W	y	$T\Sigma\Delta I/y$ (ft.-c/cone)	$T\Sigma\Delta I$ (ft.-c)
1	176	0.016	2.78
5	196	.010	1.92
10	260	.007	1.85
20	297	.005	1.54
50	367	.005	1.78
100	371	.004	1.68
218	459	.003	1.53
1082	543	.002	1.29
2610	800	.002	1.69
3410	908	.002	1.83
4050	983	.002	1.99

ΔI appreciably. The average value of ΔI per cone is given in column 3 of Table III. These values are the quotient of total summated ΔI and the number of cones acting.

DISCUSSION

The values of the summated average temporal gradient, $\sum G_{kt}$, experienced by those retinal cones which contribute to the perception of the various rectangular test objects at the threshold of visibility are essentially constant. If a dynamic condition is assumed, the values of $\sum G_{kt}$ can be considered as a measure of the stimulus, in which case the values of $\sum G_{kt}$ would be expected to be constant, since they all represent threshold stimuli. On the other hand, if a static condition is assumed, the values of $T\sum\Delta I$ can be considered as measures of the stimuli. The values as given in Table III are not constant at the threshold of visibility for the various test objects, as would be expected if the static condition of the eye existed during fixation. In addition, the average values of ΔI per cone, in foot-candles per cone, appear to be very low. If the results reported by Hartline[4] concerning the mode of functioning of the cones are valid when applied to the human eye, it follows that when the retina is subjected to a steady illuminance (constant with respect to time), although not necessarily spatially uniform, only 20 percent of the cones are generating neural responses. This is the condition which exists during *perfect* fixation. The remaining 80 percent of the cones become active, by the generation of neural responses, *only* when the incident illuminance is changing with time. The 20 percent of the cones which respond continuously to a steady illuminance *also* give a vigorous response when the illuminance is *varying* more or less rapidly with time. Hence, if our assumption of physiological nystagmus is correct, *all* the cones which are subjected to a temporal illuminance gradient (greater than some threshold value) generate a neural response. The dynamic concept, therefore, not only leads to a constant value of the stimulus at the threshold of visibility, but it also leads to a system which would be more efficient than that resulting from the assumption of a static condition during fixation.

It is difficult for us to believe that the retina and its associated nervous system are so organized that only 20 percent of the cones are active and therefore contribute to the perception of fine detail and of non-uniformity of luminance in the visual field.

While we have assumed definite values, as reported by Adler and Fliegelman,[2] for the frequency and amplitude of the physiological nystagmic motion, it appears that the *constancy* of the summated average temporal gradient does not depend upon these exact values, but rather upon the application of the principles which apply to a dynamic system of vision. This dynamic concept of the perception of non-uniformity in the visual field offers a possible explanation for the results recently reported by Fry[5] and by Lamar, Hecht, Schlaer, and Hend-

[4] H. K. Hartline, "The nerve messages in the fibers of the visual pathway," J. Opt. Soc. Am. **30**, 239 (1940).

[5] G. A. Fry, "Relation of the configuration of a brightness contrast border to its visibility," J. Opt. Soc. Am. **37**, 166 (1947).

ley.[6] For example, as reported by Lamar and co-workers, a long narrow test object would not be as efficient in producing a response as a square having the same area, since a large part of the image of the long narrow test object would produce only 64 percent as much temporal gradient as the image of the square.

Lamar and co-workers[6] determined the threshold contrast for rectangular targets which had length-width ratios varying between 2 and 200. They report that targets which have both dimensions greater than 2 minutes become increasingly less efficient as they increase in size, and state (reference 6, page 545): "For these large areas it appears that the visually critical region of the target is the ribbon just inside its perimeter and 1 or 1.5 minutes wide." This result would be expected since that area more than 1 to 1.5 minutes from the geometrical edge of the image of the test object would make no contribution to the temporal gradient on the cones of the eye. It should be noted that the illuminance gradient on the fovea would extend on both sides of the geometrical edge of the image, and that this 1 to 1.5 minutes on each side of this geometrical edge is in fair agreement with the 2.4 minutes assumed for the amplitude of the physiological nystagmic motion. The concept of a dynamic system of vision is also in agreement with the statement. "Evidently the eye makes its judgment of contrast not over the area of a target, but across its boundary."[6]

We realize that the calculations presented in this paper are based upon many assumptions, some of which may not prove to be entirely tenable. However, it appears that the concept of a dynamic system of vision does offer a possible theoretical explanation for the perception of non-uniformity of luminance in the visual field.

[6] E. S. Lamar, S. Hecht, S. Schlaer, and C. D. Hendley, "Size, shape, and contrast in detection of targets by daylight vision. I. Data and analytical description," J. Opt. Soc. Am. **37**, 531 (1947).

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 38, NUMBER 4 APRIL, 1948

The Color of Unstained Wood

PARRY MOON
Massachusetts Institute of Technology, Cambridge, Massachusetts
AND
DOMINA EBERLE SPENCER
Brown University, Providence, Rhode Island

(Received October 17, 1947)

Reflectances and colors are given for various natural woods. The results are summarized in a table, which gives the coefficients for empirical equations that fit all the spectral reflectance curves, and a second table which gives integrated values obtained by the analytic method. Clear varnish is found to reduce reflectance to approximately 85 percent of that for the unvarnished wood.

INTRODUCTION

A NOTEWORTHY modern trend in interior decoration is the use of materials having high reflectances, including unstained woods, for furniture and wall panels. Recent theoretical research[1] shows how this trend correlates with good lighting. The use of high reflectances lowers contrasts and tends to produce better visual conditions, while at the same time more usable light is obtained from a given lighting system.

A previous investigation[2] gave the reflectances

[1] Parry Moon and D. E. Spencer, *Lighting Design* (Addison-Wesley Press, Cambridge, Massachusetts, 1948); "Light distributions in rooms," J. Frank. Inst. **242**, 111 (1946); "Lighting design by the interflection method," J. Frank. Inst. **242**, 465 (1946); "A simple criterion for quality in lighting," Illum. Eng. **42**, 325 (1947).

[2] Parry Moon, "Optical reflection factors of acoustical materials," J. Opt. Soc. Am. **31**, 317 (1941); "Colors of ceramic tiles," J. Opt. Soc. Am. **31**, 482 (1941); "Wall materials and lighting," **31**, 723 (1941); "Reflection factors of floor materials," J. Opt. Soc. Am. **32**, 238 (1942); "Reflection factors of some materials used in school rooms," J. Opt. Soc. Am. **32**, 243 (1942); "Colors of furniture," J. Opt. Soc. Am. **32**, 293 (1942).

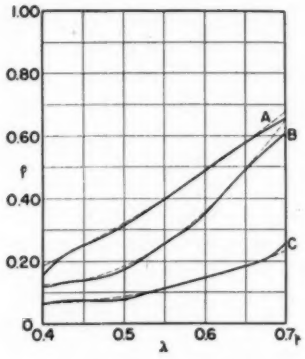


FIG. 1. Spectral reflectance curves for unvarnished woods: *A*—Faux satine; *B*—Mahogany; *C*—Brazilian rosewood; —— Experimental curve; - - - Analytic approximation.

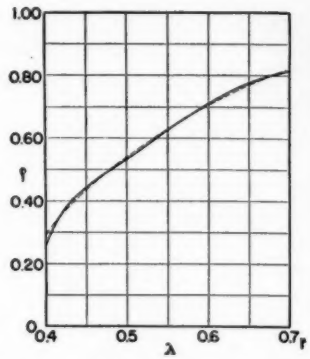


FIG. 3. Spectral reflectance curve for white harewood (unstained and unvarnished).

and colors of 600 materials used in rooms, including some stained and varnished woods. In view of the trend toward unstained woods, it was felt that a study of their reflecting possibilities would be of interest. Accordingly, spectral reflectance curves were obtained on 35 kinds of varnished and unvarnished woods. These results should indicate the maximum reflectances that can be obtained from unpigmented woods and should be of value to designers and illuminating engineers in their study of proposed lighting systems.

REFLECTANCE AND COLOR

Reflectance curves were obtained for the seventy samples by means of the Hardy spectrophotometer.[3] Three typical curves are shown in Fig. 1. They can be represented by the equation[4]

$$\rho = K_0 + K_1\lambda + K_2\lambda^2. \tag{1}$$

The heavy curves of Fig. 1 were obtained by experiment, while the dotted curves represent Eq. (1) with the constants given in Table I. These constants K_0, K_1, and K_2 were determined by making the analytical representation coincide with the experimental curve at $\lambda = 0.45$, 0.55, and 0.65μ. About two-thirds of the samples were fitted by this simple equation. A few of the highly reflecting woods were best represented by straight lines. An example is shown in Fig. 2. The equation is

$$\rho = K_0 + K_1\lambda. \tag{2}$$

Four kinds of wood had reflectance curves that were concave downward. An example is shown in Fig. 3. The equation is

$$\rho = K_{-1}/\lambda + K_0. \tag{3}$$

As shown previously,[5] analytic integration can be used to obtain the total reflectance and the colorimetric specification. Once the analytic representation of the reflectance curve has been

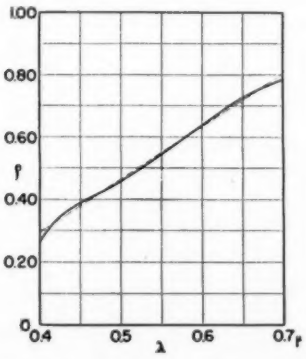


FIG. 2. Spectral reflectance curve for plain maple (unstained and unvarnished): —— Experimental curve; - - - Analytic approximation.

[3] We wish to acknowledge the assistance of the Color Measurements Laboratory of the Massachusetts Institute of Technology and to thank the staff, whose kindness made possible this investigation.

[4] Parry Moon and D. E. Spencer, "Polynomial representation of reflectance curves," J. Opt. Soc. Am. **35**, 597 (1945).

[5] Parry Moon and D. E. Spencer, "Analytical representation of trichromatic data," J. Opt. Soc. Am. **35**, 399 (1945); "Analytic expressions in photometry and colorimetry," J. Math. Phys. **25**, 111 (1946).

TABLE I.

No	Name of wood	K_{-1}	K_0	K_1	K_2
1	White harewood	−0.4899	+1.5208		
2	Avodire	−0.5418	+1.5440		
3	Plain maple		−0.3805	+1.690	
4	Ribbon prima vera	−0.5777	+1.5887		
5	Pine		−0.6420	+2.120	
6	Figured aspen		−0.5795	+1.910	
7	Figured capomo		−0.6970	+2.10	
8	Flat cut prima vera		−0.5083	+1.590	+0.3000
9	Faux satine		−0.0031	−0.1800	+1.650
10	Quartered oak		+0.2595	−1.330	+2.800
11	Gray harewood	−0.2165	+0.7710		
12	Flat cut bella rosa		+0.1462	−1.360	+3.100
13	Flat cut oak		+0.3609	−1.730	+3.050
14	Birch		+0.6730	−2.710	+3.800
15	Satinwood		−0.2522	+0.340	+1.300
16	Zebra wood		+0.2260	−1.250	+2.600
17	Cypress		+0.4332	−2.070	+3.300
18	Mahogany		+1.0931	−4.800	+5.950
19	Lacewood		+0.7457	−3.250	+4.300
20	Figured red gum		+0.3457	−1.430	+2.300
21	English oak		+0.1185	−0.870	+2.000
22	Tigerwood		+0.3665	−1.880	+3.000
23	Figured teak		+0.1876	−1.210	+2.350
24	Pearwood		+1.2077	−5.090	+5.950
25	Bubinga		+1.6546	−6.620	+7.250
26	Figured paldao		+0.4668	−1.930	+2.700
27	Flat cut walnut		+0.6377	−2.470	+3.100
28	Cherry wood		+1.2503	−5.270	+6.100
29	Claro walnut		+0.5589	−2.220	+2.850
30	Makori wood		+1.0720	−4.480	+5.200
31	Figured quartered walnut		+0.4492	−1.850	+2.500
32	Goncolo alves		+0.5916	−2.600	+3.350
33	Sapeli		+1.1838	−4.960	+5.650
34	Oriental wood		+0.0600	−0.1900	+0.600
35	Brazilian rosewood		+0.3361	−1.370	+1.750

TABLE II. Reflectance and color of unvarnished wood.

No.	Name of wood	Distribution A' x	y	ρ/ρ_0	Distribution B' x	y	ρ/ρ_0
1	White harewood	0.4793	0.4180	0.6630	0.3471	0.3585	0.6346
2	Avodire	0.4886	0.4185	0.5903	0.3589	0.3704	0.5641
3	Plain maple	0.4857	0.4134	0.5845	0.3511	0.3560	0.5589
4	Ribbon prima vera	0.4929	0.4197	0.5719	0.3655	0.3772	0.5439
5	Pine	0.4970	0.4153	0.5685	0.3673	0.3700	0.5363
6	Figured aspen	0.4971	0.4153	0.5111	0.3675	0.3702	0.4821
7	Figured capomo	0.5031	0.4163	0.5021	0.3709	0.3783	0.4702
8	Flat cut prima vera	0.4987	0.4151	0.4979	0.3696	0.3712	0.4686
9	Faux satine	0.4974	0.4124	0.4349	0.3655	0.3638	0.4092
10	Quartered oak	0.5031	0.4112	0.4177	0.3724	0.3660	0.3897
11	Gray harewood	0.4721	0.4139	0.3900	0.3353	0.3465	0.3795
12	Flat cut bella rosa	0.5182	0.4127	0.3856	0.3965	0.3845	0.3529
13	Flat cut oak	0.5054	0.4105	0.3726	0.3750	0.3663	0.3464
14	Birch	0.5000	0.4081	0.3709	0.3655	0.3560	0.3467
15	Satinwood	0.5116	0.4150	0.3680	0.3883	0.3835	0.3404
16	Zebra wood	0.5062	0.4114	0.3643	0.3769	0.3693	0.3386
17	Cypress	0.5097	0.4099	0.3327	0.3807	0.3690	0.3074
18	Mahogany	0.5238	0.4050	0.3022	0.3975	0.3705	0.2728
19	Lacewood	0.5130	0.4072	0.2991	0.3830	0.3655	0.2746
20	Figured red gum	0.4990	0.4095	0.2829	0.3653	0.3584	0.2651
21	English oak	0.5113	0.4122	0.2772	0.3852	0.3761	0.2560
22	Tigerwood	0.5158	0.4102	0.2762	0.3901	0.3755	0.2531
23	Figured teak	0.5160	0.4117	0.2668	0.3920	0.3798	0.2448
24	Pearwood	0.5238	0.4022	0.2512	0.3947	0.3633	0.2262
25	Bubinga	0.5185	0.3976	0.2505	0.3823	0.3476	0.2264
26	Figured paldao	0.5027	0.4081	0.2496	0.3692	0.3583	0.2325
27	Flat cut walnut	0.4986	0.4060	0.2432	0.3620	0.3503	0.2275
28	Cherry wood	0.5264	0.4014	0.2402	0.3980	0.3638	0.2154
29	Claro walnut	0.5013	0.4065	0.2253	0.3658	0.3535	0.2100
30	Makori wood	0.5224	0.4018	0.2180	0.3922	0.3612	0.1967
31	Figured quartered walnut	0.5037	0.4075	0.2121	0.3701	0.3580	0.1973
32	Goncolo alves	0.5176	0.4064	0.2048	0.3893	0.3680	0.1868
33	Sapeli	0.5286	0.4000	0.2032	0.4021	0.3643	0.1815
34	Oriental wood	0.4895	0.4110	0.1481	0.3540	0.3537	0.1407
35	Brazilian rosewood	0.5052	0.4063	0.1273	0.3711	0.3565	0.1181

obtained, the analytic integration takes only about five minutes per sample. The results are listed in Table II. Values of x,y, and ρ/ρ_0 are given for incident radiation from a Planckian radiator at 2842°K (distribution A')[5] and for Planckian radiation at 7000°K (distribution B'). The samples are arranged in order of decreasing reflectance. The highest reflectance, 66 percent, is obtained with white harewood, while the lowest value, 12 percent, is for Brazilian rosewood.

The values of Table II are plotted in Figs. 4 and 5. The points representing the reflected radiation are seen to cluster around the Planckian curve but with more points below it than above it. The most nearly neutral reflector is gray harewood (No. 11), while the reddest wood is sapeli (No. 33).

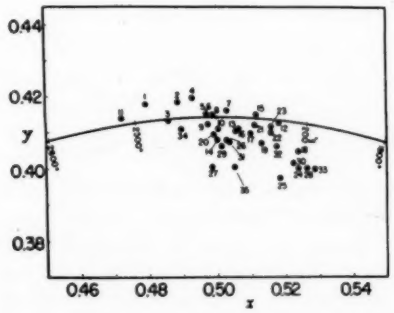


FIG. 4. Colors of 35 unvarnished woods illuminated by 2842° Planckian radiation. The numbers refer to the samples arranged in order of reflectance (Table I). The curve is the Planckian locus.

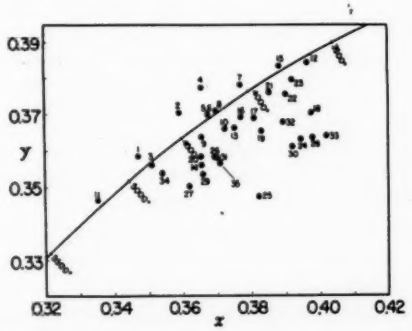


FIG. 5. Colors of the same woods, 7000° Planckian radiation.

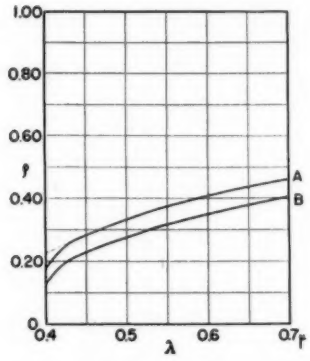


FIG. 6. Effect of varnish: *A* Gray harewood, no varnish; *B* Same material, varnished.

EFFECT OF VARNISH

The foregoing results are all for unvarnished wood and give the highest reflectances that can be obtained, at least without pigmentation. A sample of each wood was also tested when varnished. A single coat of "water-clear" varnish was used, and this varnish caused an almost constant reduction in reflectance throughout the visible spectrum. An example is shown in Fig. 6. The percentage reduction produced by the varnish differed somewhat for the different woods. On the average, however, it may be said that the reflectance with varnish is approximately 85 percent of the reflectance without varnish. If a tinted varnish had been used, the reflectance would have been reduced still further.

SUMMARY

The paper describes reflectance measurements made on 35 kinds of unvarnished wood. It is found that all samples can be represented by one of three simple equations and that analytic integrations are easily performed. The reflectances range from 0.12 to 0.66, while the colors are in all cases somewhat warmer than that of the incident radiation. Tests on similar samples that had been varnished give approximately 85 percent of the reflectance of the unvarnished samples.

The modern trend toward higher reflectances in interior decoration is in agreement with theoretical studies of room lighting.[1] Interflection theory shows that to obtain satisfactory diffusion of light for best visual conditions, one must have walls of at least 50 percent reflectance and floor and furniture of at least 30 percent reflectance.[6] The present study indicates that such reflectances can actually be obtained with unstained woods.

[6] Parry Moon and D. E. Spencer, *Lighting Design* (Addison-Wesley Press, Cambridge, Massachusetts, 1948).

Centennial Celebration of A.A.A.S.

Centennial celebrations commemorating the founding of the American Association for the Advancement of Science will be held in Washington, September 13–17, 1948. Events of the celebration, centered around the theme "One World of Science," will consist of: (1) morning symposia and panel discussions in various scientific fields of general interest; (2) afternoon excursions and visits to scientific and cultural institutions; (3) evening lectures and social affairs.

The Optical Society of America will not hold concurrent meetings with the Association at this time, but its members are invited to participate in the various events.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 38, NUMBER 4 APRIL, 1948

Electroformed Precision Prisms

BENJAMIN GOLDBERG
Engineer Research and Development Laboratories, The Engineer Center, Fort Belvoir, Virginia
(Received November 14, 1947)

This report covers an investigation for electroforming metal trihedral prisms with reflecting surfaces that approach the optical accuracy of commercial trihedral glass prisms and the field testing of the metal prisms in comparison with the glass prisms. The optical response of the best of the metal retro-directive reflectors at a distance of one mile was about three-fifths that of a commercially produced glass prism having an optically active area about three times greater than the area of the metal prism. The report concludes that the metal prism, although about as accurate as a commercially produced glass prism at a distance of one mile, is still not as accurate as its mold. However, continued improvement in the optical accuracy of the successively manufactured metal prisms indicates that the ultimate optical accuracy possible by the electroforming process has not yet been determined. Metal retro-directive reflectors produced by electroforming would cost about one-sixth as much as a commercially produced glass prism of the same optical accuracy.

I. SUBJECT

THIS report covers an investigation to determine the feasibility of producing electroformed retro-directive metal reflectors with an accuracy approaching that of a precision ground trihedral glass prism used as the mold.

II. INVESTIGATION

A retro-directive reflector is one that has the ability to redirect a beam of light back to its source with very little divergence of the returned rays. The reflectors that flash warning signals to the night driver on the highway as the headlights of his vehicle illuminate them are types of retro-directive reflectors.

A highly precise retro-directive reflector is the trihedral glass prism, commonly known as the triple reflector. This prism is pyramidal in shape with an equilaterally triangular base and three mutually perpendicular faces. A beam of light striking the base of the prism, regardless of the angle of incidence, is successively reflected from the mutually perpendicular faces and emerges parallel to the incident beam but reversed in direction and of equal intensity except for absorption and reflection losses. At present triple reflectors are made of glass and are of varying degrees of accuracy. The accuracy depends on the degree of flatness of the reflecting surfaces and the degree of perpendicularity of these surfaces to each other. Prisms with surfaces optically flat to one-quarter of a wave-length of sodium light and having angles of 90 degrees ± 1 second are expensive and difficult to make.

By use of a precision trihedral glass prism as the mold, metal replicas can be produced therefrom by the electroforming process. The reflecting surfaces of the glass prism are first given a coating of chemically deposited silver which provides the reflecting medium of the metal prism to be produced. The silver-coated glass prism is then made the cathode in a copper plating bath and copper is deposited on the silver surfaces until a thick shell of copper is electroformed thereon. Withdrawal of the glass prism from the electroformed copper shell leaves a concave, silver-surfaced replica of the reflecting surfaces of the glass prism.

A thorough investigation was made to determine the feasibility of electroforming metal prisms with reflecting surfaces that approach the optical accuracy of a trihedral glass prism used as the mold. Throughout the investigation a high precision, Mount Wilson Observatory 3-inch trihedral glass prism was used as the mold.

Four metal prisms were electroformed without the use of a preformed ring. Each of these prisms was distorted because of curling of the copper at the unsupported edges at the base of the prism.

Seven metal prisms were electroformed with the use of a preformed ring. The most difficult problem encountered in the manufacture of these metal replicas was the elimination of causes of distortion of the metal reflecting surfaces. It

TABLE I. Comparison of relative responses from metal prisms in chronological order of their manufacture.

Metal prism No.	Comparison of optical responses of metal prisms
1	Registered an optical response at the 420-yard stake but not at the 1000-yard stake
2	Registered an optical response at the 420-yard stake which was 4/3 as great as the response returned by Metal Prism 1
3	Registered an optical response two times as strong as Metal Prism 2 at the 500-yard stake but registered no response at 1000 yards
4	Registered an optical response at the 1000-yard stake but none at the 1731-yard stake
5	Registered an optical response at the 2089-yard stake which was the maximum range available
6	Registered an optical response twice as great as Metal Prism 5 at the 2089-yard range

was found that proper assembly of the mold and preformed ring prior to electroforming was of major importance in eliminating distortion of the metal prism.

On completion of the electroforming operation, the assembly (Fig. 1) consisting of the glass mold with the reflector built up on it, is placed in hot water for about thirty seconds. This expands the copper sufficiently to permit easy withdrawal of the glass mold from the replica (Fig. 2).

The last six metal prisms electroformed with preformed rings were subjected to field tests to determine the relative magnitude of response and redirecting ability when not perpendicular to the incident light, in comparison with a standard NDRC glass prism* and the Mount Wilson glass prism which was used as the mold for production of the metal prisms.

TABLE II. Relative response from Metal Prism 6 and a standard NDRC glass prism having approximately three times the optically active area.

Distance of prism from receiving equipment	Relative strength of maximum response NDRC standard glass prism	Metal prism 6
1731 yards	100 percent*	60 percent
2089 yards	100 percent*	20 percent

* No comparison is made between strength of responses at the two distances.

* The standard NDRC prism is one of many such prisms that has been produced on a mass production scale. The divergence of the returned beam from these prisms is limited to 5 seconds of arc. The divergence of the returned beam from the Mount Wilson Observatory prism is limited to 2 seconds of arc.

TABLE III. Comparison of amount of angular rotation of metal prisms and glass prisms necessary to reduce optical response to zero.

A. Azimuth Rotation

Type of prism	Amount of rotation necessary to reduce maximum response to zero: 1000-yard range	1731-yard range
NDRC (glass)	39°	30°
Prism 4 (metal)	9°	No response
Prism 5 (metal)	36°	28°
Prism 6 (metal)	35°	30°
Mount Wilson Observatory (glass)	54°	43°

B. Vertical Rotation

Type of prism	Amount of rotation necessary to reduce maximum response to zero: 1000-yard range	1731-yard range
NDRC (glass)	42°	24°
Prism 4 (metal)	20°	No response
Prism 5 (metal)	39°	26°
Prism 6 (metal)	41°	26°
Mount Wilson Observatory (glass)	60°	46°

A. Stakes were driven into the ground at five convenient locations, and the distances from the stakes to a predetermined light source were accurately measured. The stakes were at distances of 420 yards, 500 yards, 1000 yards, 1731 yards, and 2089 yards from the light source.

B. The prism to be tested was hung from the stake so that it faced the light source. The light from the source was reflected by the prism into an optical receiving unit placed almost coincident with the light source. The strength of the optical response registered by the receiving unit served as a measure of the optical accuracy of the prism.

C. Measurements were made to determine the redirecting ability of the prisms when not perpendicular to the incident rays of light. The prism was placed on an angle measuring unit and rotated horizontally and vertically.

III. RESULTS

The results of the field tests are given below:

A. *Comparison of responses of metal prisms.*—

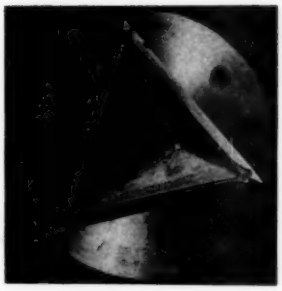


FIG. 1. Electroformed copper unit—rear view.

The relative responses of the six metal prisms, listed in chronological order of their manufacture, are shown in Table I.

B. *Comparison of responses of metal prism and glass prism.*—A comparison of the relative strength of the responses from the best of the metal prisms (Metal Prism 6) and a standard National Defense Research Committee (NDRC) glass prism with approximately three times the optically active area of the metal prism are given in Table II. (Results shown are approximate and the response from the glass prism is taken as 100 percent in each case).

C. *Comparison of angularity of metal prisms and glass prisms.*—The amount of angular rotation, both horizontally and vertically, which the best of the metal prisms was able to turn through before their responses were reduced to zero, as against similar measurements for glass prisms, is given in Table III.

IV. DISCUSSION

The results of the field tests of the electroformed metal prisms indicate a continued improvement of their optical qualities with improvement in manufacturing technique. (See Tables I and III.) This continued improvement of the metal prisms indicated that the maximum optical accuracy obtainable by electroforming has not yet been determined. Further research in manufacturing technique may result in a metal prism of higher precision.

In all of the tests performed the maximum optical response of the NDRC prism was greater than that of the metal prisms. In particular, the results given in Table II, which were obtained with Metal Prism 6 (the best of the metal prisms) and the NDRC glass prism, show the relative response of the metal prism to be three-fifths of the glass prism at 1731 yards and one-fifth of the glass prism at 2089 yards. Since optical responses of equally accurate prisms are directly proportional to their optically active areas, the difference in the maximum responses of the metal prism and the NDRC glass prism is primarily due to the fact that the optically active area of the NDRC prism is about three times as great as that of the metal prism. However, an important factor accounting for the variance in the results at 1731 yards and 2089 yards is that a minute difference in accuracy of the metal prism and the glass prism would show up more pronouncedly at a greater distance from the receiving unit. Other variables that may have affected the magnitude of the maximum responses were the limitations of the measuring instrument and the varying atmospheric conditions.

A good retro-directive reflector is able to redirect the light to its source when the reflector is held obliquely to the light source. From the test results shown in Table III, it can be seen that the angular rotation of Metal Prism 6 necessary to reduce its optical response to zero is almost equal to that required by the NDRC prism, but falls short of the arc necessary with the Mount Wilson Observatory prism.

The manufacture of a single precision 3-inch trihedral glass prism of the accuracy of the Mount Wilson Observatory prism used in this test would cost about $250. Using highly specialized production methods for producing glass

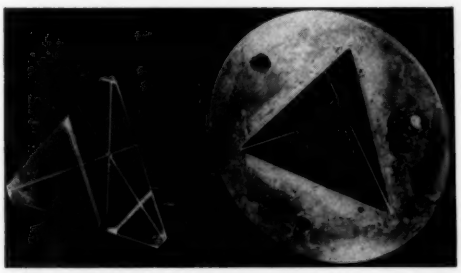


FIG. 2. Trihedral glass prism and electroformed copper replica after separation.

prisms with the accuracy of the NDRC prism, such prisms in lots of a thousand or more would cost about $50 each. The price of metal prisms in large scale production would be about $7.50 to $10.00 per prism if the following is assumed: that large quantities of molds of the Mount Wilson Observatory type can be purchased at $150 each and that a mold will produce fifty replicas before it deteriorates.

With sufficient molds high production rates may be established in the manufacture of electroformed prisms. These prisms have the following advantages and disadvantages in comparison with glass prisms produced at high production rates:

A. *Advantages.*—(1.) The molds for the electroformed prisms need not be made from high quality optical glass, whereas the glass prism must be. (2.) The price of the electroformed prisms would be about one-sixth the price of a similar glass prism. (3.) During the electroforming process, no special attention need be paid to the prisms, whereas in polishing of the glass prisms frequent checking and adjusting is required.

B. *Disadvantage.*—After separation from the mold, the metal prism cannot be further corrected, whereas the accuracy of the glass prism can at any time be improved by additional polishing.

V. CONCLUSIONS

It is concluded that:

A. The optical accuracy of an electroformed prism is largely dependent on manufacturing technique and the precision of the glass mold from which it is made.

B. Since succeeding electroformed prisms have been of an increasingly improved quality, the limit of accuracy of the metal prism has not yet been determined.

C. At a distance of one mile the angular rotation necessary to reduce the optical response to zero is about the same for the NDRC glass prism as for the electroformed prism made from the Mount Wilson Observatory prism.

D. The electroformed prism has not yet equaled the optical accuracy of its mold.

E. At present, electroformed prisms which have one-third of the optically active area of the NDRC glass prism and which register approximately three-fifths of the maximum optical response of the NDRC prism at a distance of one mile from the source, have been manufactured.

F. Electroformed metal trihedral prisms can be considered as a satisfactory substitute for commercial types of glass trihedral prisms for applications involving distances of a mile or less.

G. The manufacturing cost of the electroformed prisms is about one-sixth of that of a glass prism produced on a commercial scale.

The Behavior of an Interferometer in a Gravitational Field. II. Application to a Planetary Orbit

HERBERT E. IVES
Upper Montclair, New Jersey
(Received October 10, 1947)

The assumption of identity of properties of a body stationary in a gravitational field and of a body falling under an inverse square law of attraction, previously used to account for optical phenomena, is applied to dynamical phenomena by the extension of the same assumption to mass. Force equations corresponding to the gravitational mass

$$m_0/(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}}$$

are derived, which when solved for a planetary orbit, give the observed advance of perihelion of Mercury.

IN a paper published several years ago,[1] the so-called "interval" or "local time" as used in the general relativity gravitational theory

$$ds=dt(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}} \quad (1)$$

was derived from an assumption as to the identity of the properties of a body stationary in a gravitational field and a body falling according to an inverse square law of attraction. The optical consequences of the formula were explored, but the comment was made "There is nothing in the derivation to suggest that the formula furnishes any information outside of optics, for instance in the field of dynamics." It is the purpose of the present paper to take up the discussion where it left off, and to show how this formula may be used for the determination of a planetary orbit.

As I have shown elsewhere,[2] the "contraction factor" which transforms absolute to local time is the same factor which, appropriately placed, transforms stationary to local mass, with the result that the conservation laws (energy, momentum) hold under all conditions. In order to utilize formula (1) in dynamics the indicated procedure is therefore to use the contraction factor in its appropriate connection with *mass*. The identity of properties, when thus completely applied, states that the mass of a body moving in a gravitational field (m_g) is

$$m_g=m_0/(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}} \quad (2)$$

where m_0 is the mass in the absence of a gravitational field, the dots indicate differentiation with respect to time, and $\gamma=1-V^2/c^2$, V being the velocity at the distance r of a body falling from infinity under the force exerted by the central body which serves as the origin of coordinates for r and θ. The problem is now to determine the force equations appropriate to (2).

The force equations, in polar coordinates, where force is defined as time rate of change of momentum, are,[3] for a mass as defined by (2),

$$m_0\{\ddot{r}-r\dot{\theta}^2/(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}} +\dot{r}d/dt[1/(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}}]\}=F_R \quad (3)$$

$$m_0\{2\dot{r}\dot{\theta}+r\ddot{\theta}/(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}} +r\dot{\theta}d/dt[1/(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}}]\}=F_\theta. \quad (4)$$

As these equations stand there are too many unknowns to permit solution for the F's. We can, however, reduce the number of unknowns by reasonable assumptions. First, consider (4). If we multiply through by r and integrate we obtain

$$m_0r\cdot r\dot{\theta}/\left(\gamma-\frac{\dot{r}^2}{c^2\gamma}-\frac{r^2\dot{\theta}^2}{c^2}\right)^{\frac{1}{2}}=\int F_\theta r. \quad (5)$$

Now the left-hand side of this equation is the moment of momentum of the particle of mass (2). It is in line with the general validity of the conservation laws, above stressed, to assume that the moment of momentum shall be constant,

[1] H. E. Ives, "The behavior of an interferometer in a gravitational field," J. Opt. Soc. Am. **29**, 183 (1939).
[2] Ives, "Derivation of the Lorentz transformations," Phil. Mag. **7**, 392 (1945).
[3] A. S. Eddington, "Astronomical consequences of the electrical theory of matter," Phil. Mag. **6**, 321 (1917).

or

$$\int F_\theta r = m_0 h \tag{6}$$

and this is equivalent to putting $F_\theta = 0$, in (4).

Next, consider the value of γ or $1 - V^2/c^2$, as above defined. Under the Newtonian law of attraction $V = (2GM/r)^{\frac{1}{2}}$, giving

$$\gamma = 1 - 2GM/rc^2. \tag{7}$$

We know that the Newtonian law is deviated from only by factors of the order of v^2/c^2 or GM/rc^2, which would enter as additional terms in $1/c^4$ in (7). Consequently, we may use (7) provided we restrict ourselves to terms in $1/c^2$.

In order to find F_R, I adopt an indirect procedure, which has the merit of yielding some interesting information on other attempts at gravitational equations.

Rewrite (5) as the equivalent equation

$$m_0 r^2\dot{\theta}/(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} = m_0 h\gamma^{\frac{1}{2}}. \tag{8}$$

This is the integral of a force equation

$$m_0\{2r\dot{r}\dot{\theta} + r^2\ddot{\theta}/(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} + r^2\dot{\theta}\,d/dt[1/(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}}]\} = m_0 h\,d/dt\gamma^{\frac{1}{2}} = F_\theta' r \tag{9}$$

which refers to a local mass

$$m_0/(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}}. \tag{10}$$

Substitute in the equation $F_\theta' r = m_0 d/dt\gamma^{\frac{1}{2}}$ from (9) the value of γ from (7) and of h from (8). We get

$$F_\theta' = GMm_0/r^2(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} \cdot r\dot{r}\dot{\theta}/c^2\gamma. \tag{11}$$

From an earlier paper,[4] we recognize $r\dot{r}\dot{\theta}$ as the θ component of a force at right angles to the path of the particle, proportional to the square of the particle's velocity. The corresponding R component is $r^2\dot{\theta}^2$. We accordingly can analyze (11) as

$$F_\theta' = -GMm_0/r^2(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} \times (0 - r\dot{r}\dot{\theta}/c^2\gamma), \tag{12}$$

and write down the corresponding force in the radial direction

$$F_R' = -GMm_0/r^2(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} \times (1 + r^2\dot{\theta}^2/c^2\gamma). \tag{13}$$

[4] Ives, "The physical significance of Birkhoff's gravitational equations," Phys. Rev. **72**, 229 (1947).

The analysis of (11) chosen in deriving (12) is such as to make the masses occurring on both sides of the complete force equation formed by (9) and (12) the same, namely (10).

A point of interest appears at once. These forces, (12) and (13), will be recognized as the forces in Birkhoff's gravitational theory, as shown by a recent analysis,[4] except for the γ's occurring with the c^2 terms. These γ's merely introduce terms in GMv^2/c^4 which do not figure in the determination of orbit equations, so that for this purpose (12) and (13), which are equivalent to (3) and (4), are identical with Birkhoff's equations. The rather surprising fact brought out by the derivation of (12) and (13) is the equivalence of a multiplying factor $\gamma^{\frac{1}{2}}$ in the denominator of the local mass to a force at right angles to the motion.

Our present interest is not in the forces (12) and (13), which apply to the local mass (10), and are only a means to an end, but in the forces (3) and (4) which are in terms of local mass (2). Our next step is then to use the F_θ' and F_R' of (12) and (13) to obtain the F_θ and F_R of (3) and (4), for which purpose we solve (12) and (13) to get several useful relations. We first write out (12) and (13) at length, taking the right-hand sides as they stand, and for the left-hand sides, the left-hand sides of (3) and (4) with the substitution of local mass (10) for (2). We then multiply (12) by $\dot{r}$ and (13) by $r\dot{\theta}$ and add. The second terms on the right cancel out and we get

$$\frac{d}{dt}(\dot{r}^2 + r^2\dot{\theta}^2)/(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} - (\dot{r}^2 + r^2\dot{\theta}^2)\frac{d}{dt}(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)/(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} = -2GM\dot{r}/r^2(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} \tag{14}$$

which by multiplication and rearrangement becomes

$$\frac{d}{dt}(\dot{r}^2 + r^2\dot{\theta}^2)/(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma) = -2GM\dot{r}/r^2(1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma)$$

and upon integrating

$$r^2 + r^2\dot{\theta}^2 = (1 - \dot{r}^2/c^2\gamma^2 - r^2\dot{\theta}^2/c^2\gamma) \times (2GM/r + \cdots + a). \tag{15}$$

From this we can write, upon discarding terms in $1/c^4$,

$$1-\dot{r}^2/c^2-r^2\dot{\theta}^2/c^2=1-2GM/rc^2-a/c^2 \\ =(1-2GM/rc^2)(1-a/c^2)=\gamma/b \quad (16)$$

where b is a constant differing from unity by 10^{-8} for solar planetary velocities.

Agreeing to discard terms in $1/c^4$ as not significant in the equations of planetary orbits, we may use interchangeably, as equivalent for our purpose

$$1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma$$

and

$$1-\dot{r}^2/c^2-r^2\dot{\theta}^2/c^2.$$

To the agreed degree of approximation we can then write

$$1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma=\gamma/b \quad (17)$$

$$(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}}=\gamma/b^{\frac{1}{2}} \quad (18)$$

$$1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma \\ =(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}}/b^{\frac{1}{2}} \quad (19)$$

$$1-\dot{r}^2/c^2-r^2\dot{\theta}^2/c^2=(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}}/b^{\frac{1}{2}}. \quad (20)$$

Using these relations we rewrite (9) as

$$(2r\dot{r}\dot{\theta}+r^2\ddot{\theta})/(1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} \\ -\tfrac{1}{2}r^2\dot{\theta}d/dt(1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)/ \\ \times(1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} \\ =GM\dot{r}/r^2c^2\cdot r^2\dot{\theta}/\gamma(1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} \\ =\tfrac{1}{2}d\gamma/dt[r^2\dot{\theta}/\gamma(1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}}] \\ =\tfrac{1}{2}r^2\dot{\theta}d/dt(1-\dot{r}^2/c^2\gamma^2\cdot r^2\dot{\theta}^2/c^2\gamma)/ \\ (1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}}. \quad (21)$$

On subtracting the right-hand side from both sides, multiplying the denominators by $(1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}}$, and using relation (18), we finally get, to the chosen degree of approximation,

$$F_\theta=m_0\{(2\dot{r}\dot{\theta}+r\ddot{\theta})/(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}} \\ +r\dot{\theta}d/dt[1/(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}}]\}=0. \quad (22)$$

We could, of course, have simply noted that F_θ' was by derivation equivalent to $F_\theta=0$, but by going through the above procedure we have validated it as a method for obtaining F_R, which we now proceed to do.

Rewrite (13) as

$$(\ddot{r}-r\dot{\theta}^2)/(1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} \\ -\tfrac{1}{2}\dot{r}d/dt(1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)/ \\ (1-\dot{r}^2/c\gamma^2-r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}} \\ =-GM/r^2(1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}}(1+r^2\dot{\theta}^2/c^2\gamma).$$

From the left-hand side subtract

$$\tfrac{1}{2}\dot{r}d/dt(1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)/ \\ (1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}},$$

and from the right-hand side subtract its equivalent, from relations (17) to (21),

$$GM\dot{r}/r^2c^2\cdot\dot{r}/\gamma(1-\dot{r}^2/c^2\gamma^2-r^2\dot{\theta}^2/c^2\gamma)^{\frac{1}{2}}.$$

This gives, by the same process of multiplication and subtraction used in getting (22), and to the same degree of approximation,

$$F_R=m_0\{(\ddot{r}-r\dot{\theta}^2)/(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}} \\ +\dot{r}d/dt[1/(\gamma-\dot{r}^2/c^2\gamma-r^2\dot{\theta}^2/c^2)^{\frac{1}{2}}]\} \\ =-GMm_0/r^2(\gamma-\dot{r}^2/c^2\gamma\cdot r^2\dot{\theta}^2/c^2)^{\frac{1}{2}} \\ \times(1+r^2\dot{\theta}^2/c^2\gamma+\dot{r}^2/c^2\gamma). \quad (23)$$

These two equations, (22) and (23), are the desired force equations corresponding to the local mass (2).

As to the physical significance of these force equations: the first, as has been explained above, corresponds to a constant moment of momentum for the local mass (2). The second equation describes a Newtonian inverse square attractive force acting on the local mass (2), increased by a term proportional to the square of the velocity of the particle (divided by $c^2\gamma$).

It now remains to find what orbit these force equations give for a planet. The solution of force equations of the form of (3) and (4), when $F_\theta=0$, has been given by Eddington, in an early discussion of the effect of variable mass on gravitational equations.[3] It is, where $1/u$ is put for r,

$$\frac{d^2u}{d\theta^2}+u=-\frac{F_R}{m_0h^2u^2\left(\gamma-\dfrac{\dot{r}^2}{c^2\gamma}-\dfrac{r^2\dot{\theta}^2}{c^2}\right)^{\frac{1}{2}}}. \quad (24)$$

Using the relations (17) and (18) and the value of γ, this gives, to terms in $1/c^2$,

$$\frac{d^2u}{d\theta^2}+u\left(1-\frac{6G^2M^2b^2}{c^2h^2}\right)=\bar{K}. \quad (25)$$

Now investigation of the effect of introducing the local mass

$$m=\frac{m_0}{\left(1-\dfrac{\dot{r}^2}{c^2}-\dfrac{r^2\dot{\theta}^2}{c^2}\right)^{\frac{1}{2}}}$$

into gravitational equations by Sommerfeld and

others,[5] showed that the equation

$$\frac{d^2u}{d\theta^2}+u\left(1-\frac{G^2M^2}{c^2h^2}\right)=\bar{K}$$

resulted, and that this represents an elliptic orbit with *one sixth* the advance of perihelion that has been observed for the planet Mercury. Consequently Eq. (25) *gives the correct advance of perhelion.*

DISCUSSION

The development here presented, together with the material of the previous paper quoted, shows that *all* the phenomena of a gravitational field are comprised in the assumption that a body stationary in a gravitational field assumes the dimensions, clock rate, and mass of a body at the same point in a Newtonian field, moving with the velocity acquired in falling from infinity under the Newtonian law of attraction.[6]

Attention may be called to the fact that the equation

$$\frac{r^2\dot{\theta}}{\left(\gamma-\frac{\dot{r}^2}{c^2\gamma}-\frac{r^2\dot{\theta}^2}{c^2}\right)^{\frac{1}{2}}}=h, \qquad (26)$$

which is here used as the starting point for determining the force equations, is a relation also obtained in the general relativity treatment,[7] as well as in the treatment by Phillips,[8] who reasons from the principle of equivalence alone, without using the general principle of relativity. These treatments also call for the same value of γ as that here used. Consequently the solution by force equations, although here worked out to complete the implications of the particular postulate of equivalence previously investigated for its optical consequences, is not restricted to this postulate. It is an alternative solution to the general relativity and Phillips treatments, from the points where these treatments arrive at Eq. (26). The force equations are actually implicit in the general relativity solution, which is commonly claimed to negate the concept of force.[9]

Note added in proof: The force equation (23) can be obtained by an alternative method not using the additive procedure of the text. If we set down the quantity $(F_R'\dot{r}+F_\theta' r\dot{\theta})$ and substitute the value of

$$\frac{\frac{d}{dt}(\dot{r}^2+r^2\dot{\theta}^2)}{\left(1-\frac{\dot{r}^2}{c^2\gamma^2}-\frac{r^2\dot{\theta}^2}{c^2\gamma}\right)^{\frac{1}{2}}}$$

from it in $(F_R\dot{r}+F_\theta r\dot{\theta})$ we obtain (23) exactly. This proves at the same time that the relation

$$\left(\gamma-\frac{\dot{r}^2}{c^2\gamma}-\frac{r^2\dot{\theta}^2}{c^2}\right)^{\frac{1}{2}}=\frac{\gamma}{b^{\frac{1}{2}}}$$

used in the additive procedure but only required to be correct to terms in $1/c^2$, is also exact.

The three expressions

$$\gamma=f(GM,r)$$

$$\frac{r^2\dot{\theta}}{\left(\gamma-\frac{\dot{r}^2}{c^2\gamma}-\frac{r^2\dot{\theta}^2}{c^2}\right)^{\frac{1}{2}}}=h$$

$$\left(\gamma-\frac{\dot{r}^2}{c^2\gamma}-\frac{r^2\dot{\theta}^2}{c^2}\right)^{\frac{1}{2}}=\frac{\gamma}{k}$$

are themselves sufficient to establish an orbit equation. In the text we have required the value $\gamma=1-2GM/rc^2$ to be correct only to terms in $1/c^2$. If we use this value to derive the orbit equation from the above three expressions the resulting orbit agrees (to significant terms) with (25). This value of γ is thus, by the demands of internal consistency, also exact.

The three expressions just listed, with the exact values dictated by the above considerations, will be recognized as those used in the General Relativity treatment, arrived at by tensor analysis from a different formulation of the Principle of Equivalence.

[5] See W. H. McCrea, *Relative Physics* (Methuen and Company, Ltd., London, 1935), pp. 29–31.

[6] While this postulate of equivalence has been expressed in terms of velocity, from which the contraction factors are obtained, the primarily significant quantity is the corresponding acceleration, which is equivalent to the upward push experienced in a true gravitational field.

[7] It is obtained by combining Eqs. (38.8) and (39.62), pp. 85 and 86 of Eddington's *Mathematical Theory of Relativity* (Cambridge, 1923).

[8] Combination of Eqs. (12b) and (13), Math. Phys. 177 (1922).

[9] If one takes the equations quoted in the two preceding notes as the entire basis for developing force equations, there is no ground for favoring (22) and (23) over the simple and symmetrical equations (12) and (13), which, as noted, are essentially the equations of Birkhoff's theory.

Letters to the Editor

"Optical Staining" of Tissue

GERMAIN C. CROSSMON
Bausch & Lomb Optical Company, Rochester, New York
November 13, 1947

IT is the purpose of this note to describe the preliminary work on a new method for the coloration of protoplasm without staining. The technique is based on the use of dark-field illumination and is related to the oblique illumination method of Wright[1] as used in petrographic work. Briefly stated, the technique is as follows.

Paraffin sections of formalin "fixed" tissue are attached to perfectly clean slides by the water or desiccation method. Affixing agents such as Mayer's albumen or celloidin cannot be used since these materials scatter light. When perfectly dry, the paraffin is removed from the sections with xylol and without staining they are mounted with cover glasses in a liquid equal to or close to the refractive index of the tissue for sodium light.

As observed with dark-field illumination, tissue elements differing widely in index from the mounting medium appear white due to the refracted and reflected white light entering the microscope objective. If the difference between the element and immersion liquid is small, certain portions of their dispersion curves will approach coincidence and light of the corresponding wave-lengths will be transmitted straight through or slightly refracted oblique to the optic axis of the microscope and thus not enter the objective. The remaining light for which the object and immersion liquid differ to a greater extent in index is refracted and reflected into the objective. The eye then sees the tissue element as colored, the color being the result of the subtraction of the wave-length passing straight through or slightly refracted from white light.

[1] F. E. Wright, Am. J. Sci. **35**, 63 (1913).

Note on Ultraviolet Absorption of Zinc Oxide Phosphors

F. H. NICOLL
RCA Laboratories Division, Radio Corporation of America, Princeton, New Jersey
December 3, 1947

IN a recent note,[1] experimental measurements of the diffuse reflectivity of ZnO phosphors were presented which indicated an absorption band coinciding with the U.V. emission of the unactivated material. The presence of this absorption band was used to explain the predominant emission of U.V. at high current density.

Some further results have now been obtained which also show the high absorption of this same sample of material for the emitted U.V. A completed cathode-ray tube was made in which the ZnO phosphor (covered by an electron-permeable aluminum backing) was deposited in several separate areas, and in which the weight of phosphor deposited was different for each area.

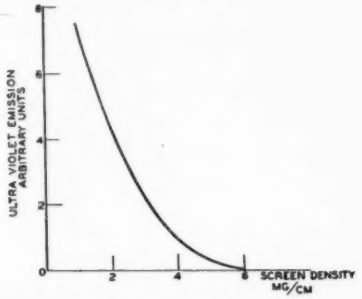


FIG. 1. Ultraviolet emission as a function of screen density for hex. ZnO:Zn phosphor under high current density bombardment 100-μa focused and scanning on a 3-in. X4-in. television pattern at 20 kv.

Figure 1 illustrates the results obtained. The U.V. output for the thinnest screen (viewed on the unbombarded side) was as much as 50 times as high as that for the thickest. At the same time, for the accelerating voltages used, the visible light output was actually greater for the thick screen because of the more efficient use of the electrons and the fact that the material absorbs very little visible light.

These results also indicate that optimum screen thickness for a given electron beam voltage is greatly affected by absorption of light in the screen. For more commonly used phosphors, optimum screen densities of several mg/sq. cm were obtained when only a small amount of light absorption was present.[2] However, in the present ZnO screen with high ultraviolet absorption, optimum screen density for U.V. emission is less than 1 mg/sq. cm at 20 kv.

[1] R. E. Shrader and H. W. Leverenz, J. Opt. Soc. Am. **37**, 939 (1947).
[2] F. H. Nicoll, *Preparation and Characteristics of Luminescent Materials* (John Wiley and Sons, Inc., New York, in course of publication).

Nomographs for Spectral Band Width of the Beckman DU Spectrophotometer[1]

H. M. HAENDLER
Department of Chemistry, University of New Hampshire, Durham, New Hampshire
December 23, 1947

ABSORPTION spectra have often been measured with variable spectral band widths, leading to some confusion and difficulty when spectra are compared. Fortunately, it is becoming more common to determine spectra at constant spectral band widths. As suggested by Gibson and Balcom[2] in their review of the Beckman DU Spectrophotometer, this is of particular importance in comparing results on different instruments or in analytical work with any instrument.

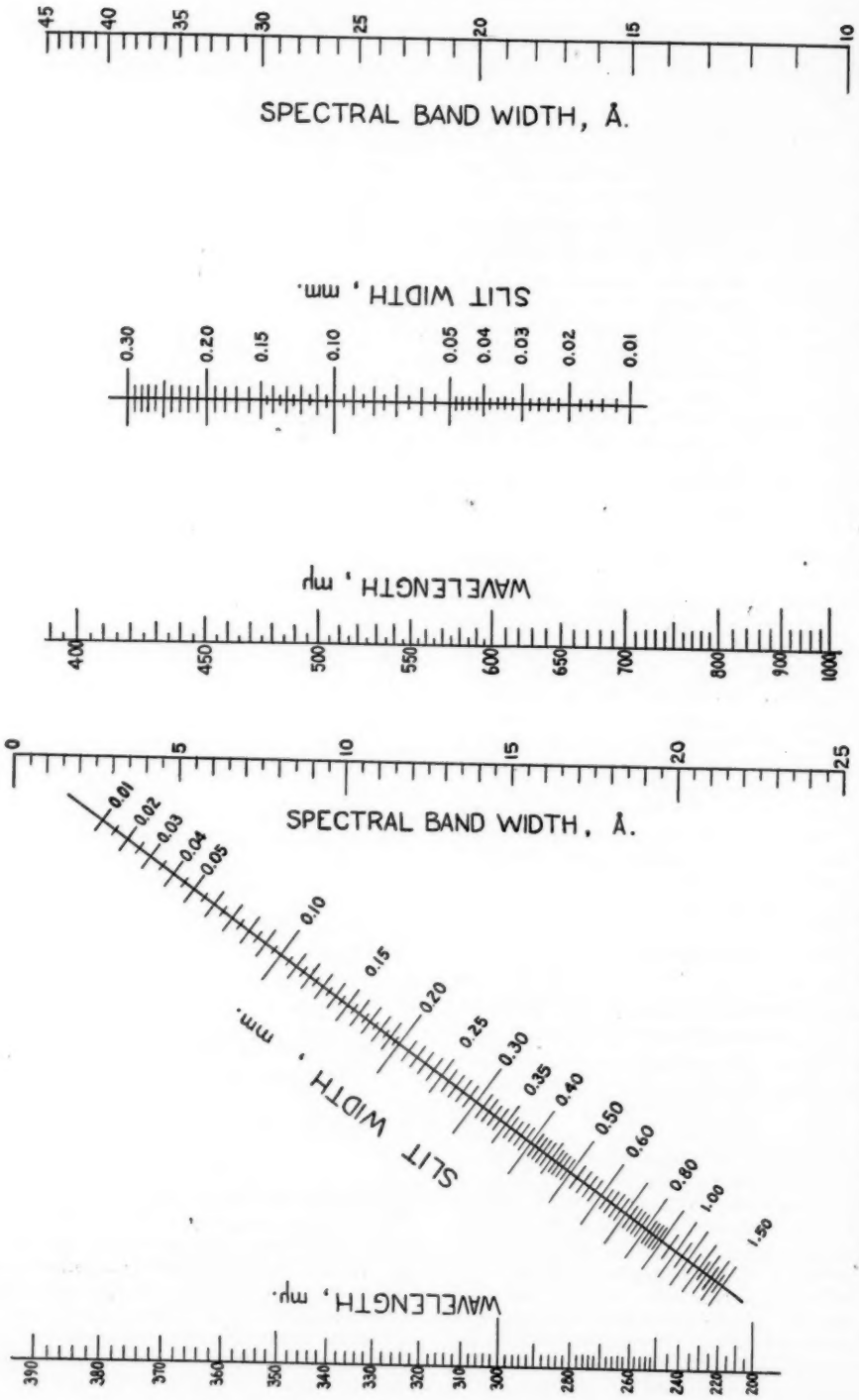


FIG. 2. Nomograph for spectral band width of the Beckman DU Spectrophotometer in the visible and low infrared range.

FIG. 1. Nomograph for spectral band width of the Beckman DU Spectrophotometer in the ultraviolet range.

The relation between spectral band width B and instrument slit width X is given for the Beckman instrument as

$$X=(B-0.02W)/W,$$

where W is the spectral band width per mm of slit width. The values of W at various wave-lengths are given by the manufacturer (Chart DD 25). The formula corrects for aberrations which widen the spectral band, as discussed by Cary and Beckman.[3]

To facilitate these calculations, and to avoid constant reference to the dispersion data, two nomographs have been prepared for use with the Beckman DU Spectrophotometer. Figure 1 gives the slit width settings in mm, for variable spectral band widths in A, for the wave-lengths from 200 to 390 mμ. Figure 2 covers the range of wave-lengths from 390 to 1000 mμ. A straight-edge, preferably transparent, connecting wave-length and desired spectral band width intersects the center scale at the proper slit setting.

[1] This work was part of a research program supported by a Research Corporation Grant-in-Aid.
[2] K. S. Gibson and M. M. Balcom, J. Research Nat. Bur. of Stand. **38**, 601 (1947).
[3] H. H. Cary and A. O. Beckman, J. Opt. Soc. Am. **31**, 682 (1941).

Electro- and Magneto-Optical Shutters for Aerial Cameras

ROY C. GUNTER, JR.
Boston University, Boston, Massachusetts
January 13, 1948

IT is felt in view of recent inquiries that certain phases of an investigation on the applicability of electro-optical and magneto-optical effects as shutters for aerial cameras might be of interest to others.

The first approach was the utilization of the Kerr effect in nitrobenzene. Nitrobenzene has been successfully used in this manner by many people, among them Anderson[1] in his researches on the velocity of light. It might be worth while to emphasize here the absolute necessity of using chemically *pure* nitrobenzene. Lack of this precaution usually results in a minor explosion. This method was given up because of the aperture required for the camera under study. Although it is possible to get very wide aperture Kerr cells using two comb-type electrodes, it would have been necessary to use voltages on the order of 500,000 volts. For aircraft installation this was felt undesirable.

Magneto-optic effects exhibited by certain colloids had been extensively investigated by Dr. Hans Mueller of the Massachusetts Institute of Technology[2] and others. Consultation with Dr. Mueller substantiated evidence that these effects were, in general, very slow acting.

The method finally adopted in December of 1946 employed the electro-optic effect to be found in ammonium di-hydrogen phosphate. The magnitude of the effect was large and had originally been investigated by Busch.[3] While light could be controlled in this manner, there were several serious problems to be considered:

1. Could the crystals be assembled into large sheets by some sort of mosaic process?
2. Could the crystals be pulsed so as to give rise and decay times negligible with respect to the total open time?
3. What angular aperture might reasonably be expected?
4. Would the crystal, when open, be a limiting factor in the resolution that might be expected from the optical system in the air?
5. Would a shutter made in this manner be sufficiently stable mechanically, electrically, and thermally for aerial use?
6. Would the percent transmission be acceptable?

Invaluable advice on these problems was given by Dr. Hans Jaffee and Mr. George Gotschall of the Brush Development Company, Cleveland, Ohio. At least partial answers to the above questions were available from these men as well as Dr. H. Mueller and Dr. Walter G. Cady,[4] and the literature.

With reference to the preceding questions, the answers were:

1a. Yes—ADP (code name PN) crystals had been successfully made into a mosaic in order to produce a sound generator of large area, but in the assembly no consideration had had to be given to its optical characteristics.

2a. Probably not—trying to pulse the crystals was, in the words of one reference, "like hitting it with a hammer."

3a. Between 3° and 6° depending upon:
(a) strength of electric field,
(b) the effective percentage of radial decrease in density that could be tolerated,
(c) the exposure and type of film.

4a. Virtually no definite information but general consensus was that there would be some loss in resolving power but probably not too much.

5a. Information disagreed—but general agreement was that the crystals were delicate thermally, and mechanically. Electrically they behaved as would similar dielectrics.

6a. Probably not—when all losses were taken into consideration the percent transmission would be about 25 percent. This figure is a far cry from the 100 percent available with mechanical shutters. The possibility of synchronizing several camera shutters with an accuracy in the low microsecond range and with virtually no vibration induced in the cameras, however, was deemed sufficient to warrant further study.

For experimental purposes the optical line-up consisted of a point source (0.003-in. diameter), a collimating lens, the crystal assembly, a 931A photo-multiplier with adjustable load resistor, and a Dumont 164E oscilloscope. The source of voltage pulses was an electronic circuit consisting of a Dumont Model 185A square wave generator, a 6SN7 voltage amplifier, an 807 power amplifier, and a 304TH output tube operating with a 20,000-ohm load resistor. The

crystal was in parallel with the load resistor. The 304TH was normally off and was turned on and off by the square waves from the 807. The *B* value was 5000 volts. This circuit is admittedly not ideal but was satisfactory for preliminary tests. Theory predicts that the maximum "opening" of the shutter will occur with approximately 9000 volts *across the crystal*. The 5000 volts was chosen largely because a power supply of this size was available, and it was felt sufficient to give the desired answers.

It will be noted that the words "across the crystal" have been italicized. The reason for this lies in the difficulty of doing just that—getting the voltage *across the crystal*. If the field were applied perpendicular to the optical axis of the system, the problem would be simple. In this system, however, the field must be applied parallel to the optical axis. This fact, while not a problem in a sonic generator, poses for light valves the problem of creating an electrode which: (a) is transparent, (b) does not affect the crystal, (c) is stable electrically, mechanically, and thermally, and (d) is capable of flexing with the crystal.

The first electrode used was constructed by laminating the crystal between two Nesa-coated plates (obtained from the Pittsburgh Plate Glass Company). Preliminary tests have not been successful in achieving uniform electrical contact between the Nesa and the crystal, resulting in the major part of the applied voltage appearing across the air gaps between the Nesa plates and the crystal. This indicated the need of some transparent conducting cement which met the above-mentioned specifications. One cement that has been used with moderate success is composed of sheet gelatine dissolved in glycerine with phosphoric acid added and desiccated until the resistance between two cylindrical test electrodes (approximately $\frac{1}{16}$" in diameter) separated by $\frac{1}{2}$" and immersed to a depth of $\frac{3}{4}$" measured about 2000 ohms. With these techniques, supplemented by a camera using type B Pan emulsion film, the following information was gathered and is presented as possibly an aid to others working in similar fields (see items 1–6 and 1b–6b):

1b. No work has been done on this because other problems were deemed to be more important.

2b. The crystals can be pulsed. Our experiments showed that it is possible to raise the voltage across the crystal at least 5000 volts in less than 10 microseconds. No attempt was made at shaping this pulse to get a faster rise time because it was already at least ten times better than necessary for an aerial camera shutter. Inspection of the light transmitted through the crystal showed a rise and fall time very close to this value. The theoretical voltage rise time may be calculated from the measured capacitance of 50 mmf/in.[3] for crystals $\frac{1}{16}$" thick and the resistance in the circuit.

3b. The angular aperture measured approximately 15°. In assessing this figure, *very* careful attention should be given to 3a (b) and 3a (c).

4b. Initial experiments indicated that the resolution of the crystal was satisfactory. More careful checks showed that this indication was correct: no loss of resolution was observed through a system that attained 45 lines per mm both with and without the crystal.

5b. Our experience was that the crystals are extremely stable, mechanically, thermally, and electrically. They, of course, must be handled with the same care as would be exercised in handling any optical surface. Skin moisture will affect ADP, however, to a greater extent than it will the usual glasses.

6b. Information in 6a, accumulated by Dr. Hans Jaffee and Mr. George Gotschall, was substantially verified. The exposure ratio between open and closed positions can be improved by using type H Polaroid of high general dichroic ratio and specially selected to have a very high degree of polarization in the red end of the spectrum up to 760 millimicrons. The light below 460 is cut out by a Wratten K-2 filter. Our preliminary experiments indicate that this exposure ratio is approximately 1:60.

Detailed information on further chemical, physical, mechanical and electro-optical characteristics of PN may be obtained by reference to three technical papers available from the Brush Development Company and entitled:

1. "Data on piezoelectric PN crystals,"
2. (a) "Piezoelectric studies of primary phosphates and arsenates,"
   (b) "The order of magnitude of piezoelectric crystals,"
3. "Basic design constants for piezoelectric crystals."

The Boston University Optical Research Laboratory wishes to express its appreciation to the Physics Department of Clark University for the use of certain of its facilities and the advice of its personnel in connection with part of the research on this project. An airborne shutter is currently being designed on the basis of the above data; details of this design and data accumulated from in-the-air tests will be published as soon as possible.

The author wishes to thank Mr. Richard F. Atwood for his aid and many contributions throughout the course of the experiments.

[1] W. C. Anderson, "A measurement of the velocity of light," Rev. Sci. Inst. **8**, 239 (1937).
[2] H. Mueller and M. H. Samos, "Magneto optical properties of ferromagnetic suspensions," Phys. Rev. **61**, 631 (1942).
[3] G. Busch, "New Seignette-electrics," Helv. Phys. Acta **2**, 269 (1938). G. Busch and E. Ganz, "Dielectric measurements on KH_2PO_4 and KH_2AsO_4 at low temperatures," Helv. Phys. Acta **15**, 501 (1942).
[4] W. G. Cady. Dr. Cady is Professor of Physics, Emeritus, at Wesleyan University and author of *Piezoelectricity* (McGraw-Hill Book Company, Inc., New York, 1946).